scanned

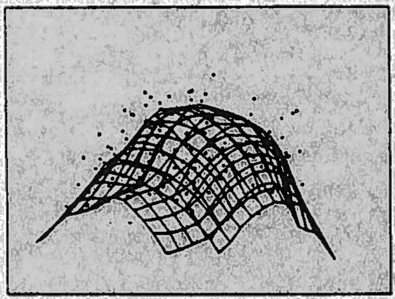
Department Copy

BETTER BOOTSTRAP CONFIDENCE INTERVALS

Bradley Efron

Technical Report No. 14
November 1984

Laboratory for Computational Statistics



Department of Statistics
Stanford University

BETTER BOOTSTRAP CONFIDENCE INTERVALS*

Bradley Efron

Department of Statistics

and

Stanford Linear Accelerator Center

Stanford, California

LCS Technical Report No. 14

November 1984

ABSTRACT

We consider the problem of setting approximate confidence intervals for a single parameter θ in a multiparameter family. The standard approximate intervals based on maximum likelihood theory, $\hat{\theta} \pm \hat{\sigma} z^{(\alpha)}$, can be quite misleading so, in practice, tricks based on transformations, bias, corrections, etc., are often used to improve their accuracy. The bootstrap confidence intervals discussed in this paper automatically incorporate such tricks without requiring the statistician to think them through for each new application, at the price of a considerable increase in computational effort. In addition to parametric families, bootstrap intervals are also developed for nonparametric situations.

* Work supported by a NSF Mathematical Sciences Postdoctoral Research Fellowship, National Foundation Grant MCS80-24649 Office of Naval Research contract N00014-83-K-0472 and grant N00014-83-G-0121, and U.S. Army Research Office under contract DAAG29-82-K-0056

Better Bootstrap Confidence Intervals

Bradley Efron

1. Introduction.

This paper concerns setting approximate confidence intervals for a real-valued parameter θ , in a multi-parameter family. The nonparametric case, where the number of nuisance parameters if infinite, is also considered. The word "approximate" is important because in only a few special situations can exact confidence intervals be constructed. Table 1 shows one such situation: the data (y_1,y_2) is bivariate normal with unknown mean vector (η_1,η_2) , covariance matrix the identity; the parameters of interest are $\theta = \eta_2/\eta_1$, and also $\phi = 1/\theta$. Fieller's construction (1954) gives central 90% interval (5% error in each tail) of [.29,.76] for θ , having observed $\underset{\sim}{y} = (8,4)$. The corresponding interval for $\phi = 1/\theta$ is the obvious mapping $\phi \in [1/.76,1/.29]$.

	for θ	(R/L)	for ϕ	(R/L)
Exact Interval	[.29,.76]	(1.21)	[1.32,3.50]	(2.20)
Standard Approximation (1.1)	[.27,.73]	(1.00)	[1.08,2.92]	(1.00)
MLE	$\hat{\theta} = .5$		$\hat{\phi} = 2$	

Table 1. Central 90% confidence intervals for $\theta = \eta_2/\eta_1$ and for $\phi = 1/\theta$, having observed $(y_1,y_2) = (8,4)$ from a bivariate normal distribution $\underset{\sim}{y} \sim N_2(\underset{\sim}{\eta},\underset{\sim}{I})$. The exact intervals are based on Fieller's construction. R/L = ratio of right side of interval, measured from the MLE, to the left side. The exact intervals are markedly asymmetric.

Table 1 also shows the standard approximate intervals

$$\theta \in [\hat{\theta}+\hat{\sigma}z^{(\alpha)}, \hat{\theta}+\hat{\sigma}z^{(1-\alpha)}] , \tag{1.1}$$

where $\hat{\theta}$ is the maximum likelihood estimate (MLE) of θ, $\hat{\sigma}$ is an estimate of its standard deviation, often based on derivatives of the log likelihood function, and $z^{(\alpha)}$ is the $100\cdot\alpha$ percentile point of a standard normal variate. In Table 1, $\alpha = .05$, and $z^{(\alpha)} = -z^{(1-\alpha)} = -1.645$.

The standard intervals (1.1) are extremely useful in statistical practice because they can be applied in an automatic way to almost any parametric situation. However they can be far from perfect, as the results for ϕ show. Not only is the standard interval for ϕ quite different from the exact interval, it is not even the obvious transformation $[1/.73, 1/.27]$ of the standard interval for θ.

Approximate confidence intervals based on bootstrap computations were introduced by the author (1981,1982). Like the standard intervals, these can be applied automatically to almost any situation, though at greater computational expense than (1.1). Unlike (1.1), the bootstrap intervals transform correctly, so for example the interval for $\phi = 1/\theta$ is obtained by inverting the endpoints of the interval for θ. They also tend to be more accurate than the standard intervals. In the situation of Table 1, the bootstrap intervals agree with the exact intervals to three decimal places. Efron (1984) shows that this is no accident; there is a wide class of problems for which the bootstrap intervals are an order of magnitude more accurate than the standard intervals.

In those problems where exact confidence limits exist, the endpoints are typically of the form

$$\hat{\theta} + \hat{\sigma}\left(z^{(\alpha)} + \frac{A_n^{(\alpha)}}{\sqrt{n}} + \frac{B_n^{(\alpha)}}{n} + \ldots\right) , \qquad (1.2)$$

where n is the sample size. The standard intervals (1.1) are <u>first order correct</u> in the sense that the term $\hat{\theta}+\hat{\sigma}z^{(\alpha)}$ asymptotically dominates (1.2). However the second order term $\hat{\sigma}A_n^{(\alpha)}/\sqrt{n}$ can have a major effect in small sample situations.

It is this term which causes the asymmetry of the exact intervals about the MLE, as seen in Table 1. As a point of comparison, the student-t effect is of third order magnitude, comparable to $\hat{\sigma}B_n^{(\alpha)}/n$ in (1.2). The bootstrap method described in Efron (1984) was shown to be <u>second order correct</u> in a certain class of problems, automatically producing intervals of correct second order asymptotic form $\hat{\theta} + \hat{\sigma}(z^{(\alpha)} + \frac{A_n^{(\alpha)}}{\sqrt{n}} + \ldots)$.

This paper describes an improved bootstrap method which is second order correct in a wider class of problems. This wider class includes all the familiar parametric examples where there are no nuisance parameters, and where the data has been reduced to a one-dimensional summary statistic, with asymptotic properties of the usual MLE form. (Section 4).

The bootstrap methods described here apply to either parametric or nonparametric situations. We will begin with the simplest parametric situations and work toward the full nonparametric case near the middle of the paper. In order to get the main ideas across, some important technical points are deferred to the later Sections.

2. Bootstrap Confidence Intervals.

This section describes the construction of improved bootstrap confidence intervals. They are obtained by a simple modification of the <u>bias-corrected bootstrap method</u>, (BC Method), introduced in Efron (1981,1982). First we describe the BC method.

Let $\underset{\sim}{y}$ represent all the available data, and suppose that $\underset{\sim}{y}$ is drawn from an unknown probability distribution P , belonging to a known family of distributions $\mathcal{P}$. A familiar example is where $\underset{\sim}{y} = (x_1, x_2, \ldots, x_n)$, the x_i being obtained by independent identical draws (i.i.d.) from a distribution $F_{\underset{\sim}{\eta}}$ which belongs to a family $\mathcal{F}$ indexed by an unknown parameter vector $\underset{\sim}{\eta}$.

There is a real-valued parameter $\theta = t(P)$ for which we have a point estimate, say $\hat{\theta} = s(\underset{\sim}{y})$, but further desire a confidence interval. All bootstrap methods begin as follows: Having observed $\underset{\sim}{y}$, we first estimate P by some estimation rule $\hat{P} = \hat{P}(\underset{\sim}{y})$. In the i.i.d. case for instance, we might estimate $\underset{\sim}{\eta}$ by its MLE $\underset{\sim}{\hat{\eta}}$, and then take $\hat{P} = F^n_{\hat{\eta}}$, the exponent indicating n independent draws of x_i from $F_{\hat{\eta}}$. In the nonparametric situation, where $\underset{\sim}{y} = (x_1, x_2, \ldots, x_n)$ but the i.i.d. observations x_i can come from any distribution F on their sample space, we would usually take $\hat{P} = \hat{F}^n$, where $\hat{F}$ is the empirical distribution putting mass $1/n$ on each observed value x_i. This is the original context in which the bootstrap was suggested. However in most of this paper, except for Sections 6 and 7, we will be working with the parametric bootstrap.

The BC method produces an approximate $1-2\alpha$ central confidence interval for θ by resampling from $\hat{P}$:

(i) Independent bootstrap data sets $\underset{\sim}{y}^*_1, \underset{\sim}{y}^*_2, \ldots, \underset{\sim}{y}^*_B$ are drawn from $\hat{P}$. (The number of resamples for confidence interval construction tends to be large, on the order of $B = 1000$, see Section 8. We will assume in what follows that B is very large, so that fluctuations in the bootstrap intervals due to small B are eliminated. In parametric situations B can often be made essentially infinite by using standard parametric expansions, instead of Monte Carlo sampling, to construct the bootstrap distribution, see Efron (1984, 1984A).)

(ii) For each y^*_b, $b = 1, 2, \ldots, B$, the bootstrap estimate $\hat{\theta}^*_b = s(\underset{\sim}{y}^*_b)$ is calculated.

(iii) The bootstrap cumulative distribution function (cdf) of the $\hat{\theta}^*$ values is constructed, say

$$\hat{G}(s) = \frac{\#\{\hat{\theta}^*_b < s\}}{B} \qquad -\infty < s < \infty . \tag{2.1}$$

(iv) The quantity

$$z_0 = \Phi^{-1}(\hat{G}(\hat{\theta})) \qquad (2.2)$$

is evaluated, $\Phi(z) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{z} e^{-\frac{1}{2}s^2}\, ds$, the standard normal cdf.

(v) Finally, the BC interval is defined to be

$$\theta \in [\hat{G}^{-1}(\Phi(2z_0+z^{(\alpha)})),\ \hat{G}^{-1}(\Phi(2z_0+z^{(1-\alpha)}))] \ , \qquad (2.3)$$

where $z^{(\alpha)} = \Phi^{-1}(\alpha)$ as before.

Notice that if $\hat{G}(\hat{\theta}) = .5$, that is if half of the $\hat{\theta}_b^*$ values are less than the actual estimate $\hat{\theta}$, then $z_0 = 0$, and the BC interval (2.3) is simply $\theta \in [\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)]$. In other words, we use the obvious percentiles of the bootstrap distribution of $\hat{\theta}^*$ to form an approximate confidence interval for θ. If $z_0 \neq 0$, definition (2.3) makes a bias correction, often a quite large one, as motivated in Section 10.7 of Efron (1982).

For the situation of Table 1, the BC method produces intervals agreeing very closely with the exact Fieller solution. Table 2 shows a less successful application, pointed out by N. Schenker (1983). The data is the single observation $y \sim \theta\chi_{19}^2$, and a confidence interval is desired for the scale parameter θ. In this case the BC interval based on $\hat{\theta} = y/19$ is a definite improvement over the standard interval (1.1), but goes only about half as far as it should toward achieving the asymmetry of the exact interval, $\theta \in \hat{\theta}[19/\chi_{19}^{2(1-\alpha)}, 19/\chi_{19}^{2(\alpha)}]$.

Why does the BC method work better in Table 1 than in Table 2? The main result of Efron (1984) is the following, which applies to Table 1: suppose $\underset{\sim}{y}$ is multivariate normal $\underset{\sim}{y} \sim N_k(\underset{\sim}{\eta},\underset{\sim}{I})$, $\theta = t(\underset{\sim}{\eta})$, and we estimate θ by its MLE $\hat{\theta} = t(\underset{\sim}{y})$. Then the BC method is second order correct, as defined in Section 1. More generally, if there exists multivariate transformations g and h such that $g(\underset{\sim}{y}) \sim N_k(h(\underset{\sim}{\eta}),\underset{\sim}{I})$, then the BC interval for θ based on the MLE $\hat{\theta}$ is

1.	Exact	$\hat{\theta}[.631,1.88]$	$R/L = 2.38$
2.	BC	$\hat{\theta}[.580,1.69]$	$R/L = 1.64$
3.	Standard (1.1)	$\hat{\theta}[.466,1.53]$	$R/L = 1.00$
4.	$BC_{.1077}$	$\hat{\theta}[.630,1.88]$	$R/L = 2.37$
5.	Nonparametric BC_a	$\hat{\theta}[.640,1.68]$	$R/L = 1.88$

Table 2. Central 90% confidence intervals for θ, having observed $y \sim \theta\chi^2_{19}$. The BC method, based on the MLE $\hat{\theta} = y/19$, (line 2), is only a partial improvement over the standard intervals. The improved bootstrap method described in this section (line 4) agrees almost perfectly with the exact interval. Its nonparametric version (line 5) is discussed in Section 6.

still second order correct. This last result does not require the statistician to know the normalizing transformations g and h, only that they exist (because the BC interval (2.3) is invariant under all such transformations, when $\hat{\theta}$ is the MLE).

The situation of Table 2 is $y \sim \theta\chi^2_{19}$, or equivalently $\hat{\theta} \sim \theta(\chi^2_{19}/19)$. The results of Efron (1982A) show that there is a single monotone transformation g such that, to a good approximation, $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$ satisfy

$$\hat{\phi} = \phi + \sigma_{\phi}(Z - z_0) , \qquad (Z \sim N(0,1)) \tag{2.4}$$

and

$$\sigma_{\phi} = 1 + a\phi . \tag{2.5}$$

The constants in (2.4), (2.5) are $z_0 = .1082$, $a = .1077$. See Section 9, and Remark E, Section 10. For general situations of form (2.4), (2.5), we will assume that $\phi > -1/a$ if $a > 0$, so $\sigma_{\phi} > 0$, and likewise $\phi < -1/a$ if $a < 0$. The constant a will typically be in the range $|a| \leq .2$ as will z_0. [Please note the Corrigenda to Efron (1982A).]

Notice that if a were 0 in (2.5), then (2.4) would be a normal translation family. In this case the obvious confidence interval

$\phi \in [\hat{\phi}+z_0+z^{(\alpha)}, \hat{\phi}+z_0+z^{(1-\alpha)}]$ would map into a correct confidence interval for θ, via the inverse transformation $\theta = g^{-1}(\phi)$. As a matter of fact the resulting interval for θ equals the BC interval. This is the rationale for definition (2.3). The advantage of (2.3) is that the statistician need not know the mapping g.

The improved bootstrap method of this paper consists of a recipe for handling the situation $a \neq 0$. Suppose $\hat{\theta}$ is a real-valued statistic whose density f_θ depends only on a real-valued parameter θ, and suppose also that there exists a monotone mapping g such that $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$ satisfy (2.4), (2.5).

Lemma 1. Under the conditions just stated, the correct confidence interval for θ, based on $\hat{\theta}$, is

$$\theta \in [\hat{G}^{-1}(\Phi(z[\alpha]), \hat{G}^{-1}(\Phi(z[1-\alpha])] \ , \tag{2.6}$$

where

$$z[\alpha] = z_0 + \frac{(z_0+z^{(\alpha)})}{1 - a(z_0+z^{(\alpha)})} \ , \tag{2.7}$$

and likewise for $z[1-\alpha]$. (Proof below.)

Here $\hat{G}$ is the (parametric) bootstrap cdf of $\hat{\theta}^* \sim f_{\hat{\theta}}$. If $a = 0$ then interval (2.6) is the same as (2.3), but if $a \neq 0$, different percentiles of the bootstrap distribution are employed. We will call (2.6) *the BC_a interval*. The constant a is discussed further in Sections 3 and 9.

Example: for the situation of Table 2, where $z_0 = .1082$, $a = .1077$, we calculate $z[.05] = -1.210$, $z[.95] = 2.270$, so $\Phi(z[\alpha]) = .1131$, $\Phi(z[1-\alpha]) = .9884$. The bootstrap distribution is $\hat{\theta}^* \sim \hat{\theta}(\chi^2_{19}/19)$, with quantiles $\hat{G}^{-1}(.1131) = \hat{\theta}.630$, $\hat{G}^{-1}(.9884) = \hat{\theta}\ 1.877$. This gives the $BC_{.1077}$ interval shown in line 4. [Bartlett (1953) discusses this same example. The BC_a method can be thought of as a computer-based way to numerically carry out Bartlett's program of improved approximate confidence intervals without having to do any theoretical calculations.]

The proof of the lemma begins by showing that the BC_a interval for ϕ, based on $\hat{\phi}$, is correct in a certain obvious sense: notice that (2.4), (2.5) give

$$\{1+a\hat{\phi}\} = \{1+a\phi\}\{1+a(Z-z_0)\} . \tag{2.8}$$

Taking logarithms puts the problem into standard translation form,

$$\hat{\zeta} = \zeta + W , \tag{2.9}$$

$\hat{\zeta} = \log\{1+a\hat{\phi}\}$, $\zeta = \log\{1+a\phi\}$, and $W = \log\{1+a(Z-z_0)\}$. This example is discussed more carefully in Sections 4 and 8 of Efron (1982A), where the possibility of the bracketed terms in (2.8) being negative is dealt with. Here it will cause no trouble to assume them positive so that it is permissable to take logarithms. In fact the transformation to (2.9) is only for motivational purposes. A quicker but less informative proof of the Lemma is possible working directly on the ϕ scale.

The translation problem (2.9) gives a natural central $1-2\alpha$ interval for ζ having observed $\hat{\zeta}$,

$$\zeta \in [\hat{\zeta}-w^{(1-\alpha)}, \hat{\zeta}-w^{(\alpha)}] , \tag{2.10}$$

where $w^{(\alpha)}$ is the $100\cdot\alpha$ percentile point for W, $\text{Prob}\{W<w^{(\alpha)}\} = \alpha$.

We will use the notation $\theta[\alpha]$ for the α-level endpoint of a confidence interval for a parameter θ. For example (2.10) says that $\zeta[\alpha] = \hat{\zeta}-w^{(1-\alpha)}$, $\zeta[1-\alpha] = \hat{\zeta}-w^{(\alpha)}$. The interval (2.10) can be transformed back to the ϕ scale by the inverse mappings $\hat{\phi} = (e^{\hat{\zeta}}-1)/a$, $\phi = (e^{\zeta}-1)/a$, $(Z-z_0) = (e^{W}-1)/a$. A little algebraic manipulation shows that the resulting interval for ϕ has α-level endpoint

$$\phi[\alpha] = \hat{\phi} + \sigma_{\hat{\phi}} \frac{(z_0+z^{(\alpha)})}{1 - a(z_0+z^{(\alpha)})} . \qquad (2.11)$$

The cdf of $\hat{\phi}$ according to (2.4) is $\Phi(\frac{s-\phi}{\sigma_\phi} + z_0)$, so the bootstrap cdf of $\hat{\phi}^*$, say $\hat{G}_\phi$, is $\hat{G}_\phi(s) = \Phi(\frac{s-\hat{\phi}}{\sigma_{\hat{\phi}}} + z_0)$. This has inverse $\hat{G}_\phi^{-1}(\alpha) = \hat{\phi}+\sigma_{\hat{\phi}}\{\Phi^{-1}(\alpha)-z_0\}$, which shows that $\hat{G}_\phi^{-1}(\Phi(z[\alpha]))$ equals (2.11). In other words, the BC_a interval for ϕ , based on $\hat{\phi}$, coincides with the correct interval (2.11), correct meaning in agreement with the translation problem interval (2.10).

The BC_a intervals transform in the obvious way: if $\hat{\phi} = g(\phi)$, $\phi = g(\phi)$, then the BC_a interval endpoints satisfy $\phi[\alpha] = g(\theta[\alpha])$ (because each bootstrap realization $\hat{\phi}_b^*$ equals $g(\hat{\theta}_b^*)$, so that all the percentiles of the two bootstrap distributions map in the same way. $\hat{G}_\phi^{-1}(\alpha) = g(\hat{G}^{-1}(\alpha))$.) This verifies Lemma 1: the transformations $\hat{\theta} \rightarrow \hat{\phi} \rightarrow \hat{\zeta}$ and $\theta \rightarrow \phi \rightarrow \zeta$ reduce the problem to translation form (2.9); the inverse transformations of the natural interval (2.10) for ζ produce the BC_a interval (2.6), (2.7).

3. The Acceleration Constant a.

The BC_a intervals (2.6), (2.7) require the statistician to calculate the bootstrap distribution $\hat{G}$, and also the two constants z_0 and a. The bias-correction constant z_0 equals $\Phi^{-1}(\hat{G}(\hat{\theta}))$, (2.2), and so can be computed directly from $\hat{G}$. What about a? If we need to know the transformation g leading to the normalized problem (2.4), (2.5) in which a was defined, then the BC_a method is practically useless. Fortunately there is a simple way to calculate a without knowledge of g.

Suppose first that $\hat{\theta}$ has density function f_θ depending only on the real parameter θ , as in Lemma 1. In this section we will show that a good approximation for the constant a is

$$a \doteq \frac{\text{SKEW}_{\theta=\hat{\theta}}(\dot{\ell}_\theta)}{6} \qquad (3.1)$$

where $\text{SKEW}_{\theta=\hat{\theta}}(X)$ indicates the skewness of a random variable X, $\mu_3(X)/\mu_2(X)^{3/2}$, evaluated at parameter point θ equals $\hat{\theta}$, and $\dot{\ell}_\theta$ is the score function

$$\dot{\ell}_\theta(\hat{\theta}) = \frac{\partial}{\partial\theta} \log f_\theta(\hat{\theta}) \ . \qquad (3.2)$$

Formula (3.1) allows us to calculate a in terms of the given density f_θ, without knowing g. Sections 5 and 6 discuss the computation of a in families with nuisance parameters. Section 9 gives a deeper discussion of a, and its relationship to other quantities of interest.

Example: For the situation $\hat{\theta} \sim \theta(\chi^2_{19}/19)$ in Table 2, standard χ^2 calculations give $\text{SKEW}(\dot{\ell}_\theta)/6 = [8/(19 \cdot 36)]^{\frac{1}{2}} = .1081$, which is quite close to the actual value $a = .1077$ derived in Section 9.

If we make smooth one-to-one transformations $\hat{\phi} = g(\hat{\theta})$, $\phi = h(\theta)$, then $\dot{\ell}_\phi(\hat{\phi}) = \dot{\ell}_\theta(\hat{\theta})/h'(\theta)$, and $\text{SKEW}(\dot{\ell}_\phi) = \text{SKEW}(\dot{\ell}_\theta)$. In other words, the right side of (3.1) is invariant under all mappings of this type. Suppose that for some choice of g and h, we can represent the family of distributions of $\hat{\phi}$ as

$$\hat{\phi} = \phi + \sigma_\phi q(Z) \qquad (Z \sim N(0,1)) \ , \qquad (3.3)$$

where σ_ϕ and $q(Z)$ are functions of ϕ and z, having at least one and two derivatives respectively. Situation (3.3) is called a general scaled transformation family (GSTF) in Efron (1982A).

Lemma 2. The family (3.3) has score function $\dot{\ell}_\phi(\hat{\phi})$ satisfying

$$\sigma_\phi \dot{\ell}_\phi(\hat{\phi}) \sim \left[Z + \frac{q''(Z)}{q'(Z)}\right]\left[\frac{1 + \dot{\sigma}_\phi q(Z)}{q'(Z)}\right] - \dot{\sigma}_\phi \qquad (Z \sim N(0,1)) \ . \qquad (3.4)$$

Here $\dot{\sigma}_\phi = \frac{\partial \sigma_\phi}{\partial \phi}$ and q' and q'' are the first two derivatives of q.

Before presenting the proof of Lemma 2, we note that it verifies (3.1): in situation (2.4), (2.5), where $\dot{\sigma}_\phi = a$, $q'(Z) = 1$, $q''(Z) = 0$, the distributional relationship (3.4) becomes

$$\sigma_\phi \dot{\ell}_\phi(\hat{\phi}) \sim (1-az_0)\left[Z + \frac{a}{1-az_0}(Z^2-1)\right] . \tag{3.5}$$

Let

$$\varepsilon_0 \equiv \frac{a}{1-az_0} , \tag{3.6}$$

a quantity discussed in Section 9. From the moments of $Z \sim N(0,1)$, (3.5) gives

$$\frac{\mathrm{SKEW}(\dot{\ell}_\phi)}{6} = \varepsilon_0 \frac{1+\frac{4}{3}\varepsilon_0^2}{(1+2\varepsilon_0^2)^{3/2}} . \tag{3.7}$$

We will see in Section 9 that for the usual repeated sampling situation both a and z_0 are order of magnitude $O(n^{-\frac{1}{2}})$ in the sample size n. This means that $\varepsilon_0 = a\cdot[1+O(n^{-1})]$, (3.6), and that $\mathrm{SKEW}(\dot{\ell}_\theta)/6 = \mathrm{SKEW}(\dot{\ell}_\phi)/6 = a[1+O(n^{-1})]$, (3.7), justifying approximation (3.1). The "constant" a actually depends on θ , but substituting $\theta = \hat{\theta}$ in (3.1) causes errors only at the third order level, $\hat{\sigma}B_n^{(\alpha)}/n$ in (1.2), and so doesn't effect the second order properties of the BC_a intervals.

Proof of Lemma 2: Starting from (3.3), the cdf of $\hat{\phi}$ is $\Phi(q^{-1}(\frac{\hat{\phi}-\phi}{\sigma_\phi}))$ so $\hat{\phi}$ has density $f_\phi(\hat{\phi}) = e^{-\frac{1}{2}Z_\phi^2}/(\sqrt{2\pi}\,\sigma_\phi\, q'(Z_\phi))$, where $Z_\phi \equiv q^{-1}((\hat{\phi}-\phi)/\sigma_\phi)$. This gives log likelihood function

$$\ell_\phi(\hat{\phi}) = -\frac{1}{2}Z_\phi^2 - \log(q'(Z_\phi)) - \log(\sigma_\phi) . \tag{3.8}$$

Lemma 2 follows by differentiating (3.8) with respect to ϕ , and noting that $Z_\phi \sim N(0,1)$ when sampling from (3.3) ∎

Suppose $z_0 = 0$ and $a > 0$ in (2.4), (2.5). Having observed $\hat{\phi} = 0$, and noticing $\sigma_{\hat{\phi}} = 1$, the naive interval for ϕ (which is almost the same as the standard interval (1.1)), is $\phi \in [z^{(\alpha)}, z^{(1-\alpha)}]$. However if the statistician checks the situation at the right endpoint $z^{(1-\alpha)}$, he finds that the hypothesized standard deviation of $\hat{\phi}$ has increased from 1 to $1+az^{(1-\alpha)}$. This suggests increasing the right endpoint to $z^{(1-\alpha)}(1+az^{(1-\alpha)})$. Now the hypothesized standard deviation has further increased to $1+az^{(1-\alpha)}(1+az^{(1-\alpha)})$, suggesting a still larger right endpoint, etc. Continuing in this way leads to formula (2.11). [Improving the standard interval (1.1) by recomputing $\hat{\sigma}$ at its endpoints is a useful idea. It was brought to my attention by John Tukey, who pointed out its use by Bartlett (1953), see for instance Bartlett's equation (17). Tukey's 1949 unpublished talk anticipates many of the same points.]

We will call a the acceleration constant in what follows because of its effect of constantly changing the natural units of measurement as we move along the ϕ (or θ) axis. Notice that we can write (2.5) as

$$\sigma_{\phi} = \sigma_{\phi_0}\left[1 + a\frac{\phi-\phi_0}{\sigma_{\phi_0}}\right], \tag{3.9}$$

so

$$a = \frac{d(\sigma_{\phi}/\sigma_{\phi_0})}{d(\frac{\phi-\phi_0}{\sigma_{\phi_0}})} \tag{3.10}$$

for any fixed value of ϕ_0. This shows that a is the relative change in σ_{ϕ} per unit standard deviation change in ϕ, no matter what value ϕ has.

The point $\phi_0 = 0$ is favored in definition (2.5) since we have set σ_0 equal to the convenient value 1. There is no harm in thinking of 0 as the true value of ϕ, the value actually governing the distribution of $\hat{\phi}$ in (2.4), because in theory we can always choose the transformation g so that this is the case, and also so that $\sigma_0 = 1$. (See Remark A, Section 10). The restriction $1+a\phi > 0$ in (2.5) causes no practical trouble for $|a| \leq .2$, since it is then at least 5 standard deviations to the boundary of the permissable ϕ region.

4. Second Order Correctness of the BC_a Intervals.

The standard interval (1.1) is based on taking literally the asymptotic approximation

$$\frac{\hat{\theta}-\theta}{\hat{\sigma}} \sim N(0,1) . \tag{4.1}$$

The BC method assumes that a more general approximation holds,

$$\frac{g(\hat{\theta})-g(\theta)}{\hat{\sigma}_g} \sim N(-z_0,1) , \tag{4.2}$$

for some constant z_0 and monotone transformation g, where $\hat{\sigma}_g = [\mathrm{Var}_\theta\, g(\hat{\theta})]^{\frac{1}{2}}_{\theta=\hat{\theta}}$. The BC_a method relaxes the assumptions one step further, to

$$\frac{g(\hat{\theta})-g(\theta)}{\hat{\sigma}_g} \sim N(-z_0,(1+ag(\theta))^2) . \tag{4.3}$$

The difference between (4.2) and (4.3) is greater than it seems: the hypothesized ideal transformation g in (4.2) has to be both normalizing and variance stabilizing, while in (4.3), g need by only normalizing. Efron (1982A) shows that normalization and stabilization are partially antagonistic goals in familiar families such as the Poisson and the binomial.

It is not surprising that a theory based on (4.3) is usually more accurate than a theory based on (4.1). In fact applied statisticians make frequent use of devices like those in (4.3), transformations, bias corrections, and even acceleration corrections, to improve the performance of approximation (4.1). The advantage of the BC_a method is that it automates these improvements, so that the statistician doesn't have to think them through anew for each new application.

Is it possible to go beyond (4.3), to find still further improvements over (4.1)? The answer is no, at least not in terms of second-order asymptotics. The theorem of this section states that for simple one-parameter problems the BC_a intervals coincide through second order with the exact intervals. In terms of (1.2), the BC_a intervals have the correct second-order asymptotic form $\hat{\theta} + \hat{\sigma}(z^{(\alpha)}+A_n^{(\alpha)}/\sqrt{n}+\ldots)$.

We consider the simple one-parameter problem $\hat{\theta} \sim f_\theta$, supposing that the $100\cdot\alpha$ percentile of $\hat{\theta}$ as a function of θ , say $\hat{\theta}_\theta^{(\alpha)}$, is a continuously increasing function of θ , for any fixed α. In this case the usual confidence interval construction gives an exact $1-2\alpha$ central interval for θ having observed $\hat{\theta}$, say $[\theta_{Ex}[\alpha], \theta_{Ex}[1-\alpha]]$, where $\theta_{Ex}[\alpha]$ is the value of θ satisfying $\hat{\theta}_\theta^{(1-\alpha)} = \hat{\theta}$. The exact interval in Table 2 is an example of this construction.

It isn't necessary that $\hat{\theta}$ be the MLE of θ. In (2.4) for instance $\hat{\phi}$ is not the MLE of ϕ. (The BC_a method is quite insensitive to small changes in the form of the estimator, see Remark B, Section 10.) However we will assume that $\hat{\theta}$ behaves asymptotically like the MLE in terms of the orders of magnitude of its bias, standard deviation, skewness, and kurtosis,

$$\hat{\theta}-\theta \sim (\frac{B_\theta}{n}, \frac{C_\theta}{\sqrt{n}}, \frac{D_\theta}{\sqrt{n}}, \frac{E_\theta}{n}) \ . \qquad (4.4)$$

Here n is the sample size upon which the summary statistic $\hat{\theta}$ is based; B_θ, C_θ, D_θ , and E_θ are bounded functions of θ (and of n, which is suppressed in the notation). Then (4.4) says that the bias of $\hat{\theta}$, B_θ/n , is $O(n^{-1})$, the standard deviation $C_\theta/\sqrt{n}$ is $O(n^{-\frac{1}{2}})$, skewness $O(n^{-\frac{1}{2}})$, and kurtosis $O(n^{-1})$. Higher cumulants, which are typically of order smaller than $O(n^{-1})$, will be assumed negligible in proving the results which follow. See Hougaard (1982) and DiCiccio (1984).

The asymptotics of this paper are stated relative to the size of the estimated standard error $\hat{\sigma}$ of $\hat{\theta}$, as in (1.2). It is often convenient in what follows to have $\hat{\sigma}$ be $O(1)$. This is easy to accomplish by transforming to $\hat{\phi} \equiv \sqrt{n}\, \hat{\theta}$, $\phi \equiv \sqrt{n}\, \theta$, so (4.4) becomes

$$\hat{\phi}-\phi \sim (\beta_\phi, \sigma_\phi, \gamma_\phi, \delta_\phi) \ , \qquad (4.5)$$

where $\beta_\phi = B_{\phi/\sqrt{n}}/\sqrt{n}$, $\sigma_\phi = C_{\phi/\sqrt{n}}$, $\gamma_\phi = D_{\phi/\sqrt{n}}/\sqrt{n}$, $\delta_\phi = E_{\phi/\sqrt{n}}/n$. Notice that $\beta_\phi = O(n^{-\frac{1}{2}})$ and $\dot{\beta}_\phi \equiv \frac{d\beta_\phi}{d\phi} = O(n^{-1})$, etc. We can just assume to begin with that $\hat{\theta}$ and θ are the rescaled quantities called $\hat{\phi}$ and ϕ above. Then the following orders of magnitude apply,

$$\begin{array}{cccc} O(1) & O(n^{-\frac{1}{2}}) & O(n^{-1}) & O(n^{-3/2}) \\ \hline \sigma_\theta & \dot{\sigma}_\theta, \beta_\theta, \gamma_\theta & \ddot{\sigma}_\theta, \dot{\beta}_\theta, \dot{\gamma}_\theta, \delta_\theta & \ddot{\beta}_\theta, \ddot{\gamma}_\theta, \dot{\delta}_\theta \end{array} \ . \qquad (4.6)$$

<u>Theorem 1</u>. If $\hat{\theta}$ has bias β_θ , standard error σ_θ , skewness γ_θ , and kurtosis δ_θ satisfying (4.6), then the BC_a intervals are second order correct.

The theorem states that $\theta_{BC_a}[\alpha]$, the α-endpoint of the BC_a interval, is asymptotically close to the exact endpoint,

$$\frac{\theta_{BC_a}[\alpha] - \theta_{Ex}[\alpha]}{\hat{\sigma}} = O(n^{-1}) \ . \qquad (4.7)$$

This isn't true for standard intervals (1.1) or the BC intervals (2.3). The proof of Theorem 1, which appears in Section 11, makes it clear that all three of the elements in (4.3), the transformation g , the bias-correction constant z_0 , and the acceleration constant a , make necessary corrections of $O(n^{-\frac{1}{2}})$ to the standard intervals based on (4.1).

McCullagh (1984) and Cox (1980) give an interesting approximate confidence interval for θ , having α-endpoint

$$\theta_{APP}[\alpha] = \hat{\theta} + \frac{1}{\sqrt{\hat{k}_2}} \left\{ z^{(\alpha)} + \frac{(3\hat{k}_2' + 2\hat{k}_{001}) + \hat{k}_{001} z^{(\alpha)^2}}{6\hat{k}_2^{3/2}} \right\} . \qquad (4.8)$$

Here $\hat{\theta}$ is the MLE of θ; if $k_2(\theta) = E_\theta \dot{\ell}_\theta^2$, the Fisher information, then $\hat{k}_2 = k_2(\hat{\theta})$ and $\hat{k}_2' = dk_2(\theta)/d\theta|_{\theta=\hat{\theta}}$; and $\hat{k}_{001} = (E_\theta \dddot{\ell}_\theta)_{\theta=\hat{\theta}}$. Formula (4.8) is based on higher-order asymptotic approximations to the distribution of the MLE. See also Barndorf-Nielsen (1984).

It can be shown, as indicated in Section 11, that $\theta_{BC_a}[\alpha]$ also closely matches (4.8), $(\theta_{BC_a}[\alpha]-\theta_{APP}[\alpha])/\hat{\sigma} = O(n^{-1})$. We see again that the BC_a method offers a way to avoid theoretical effort, at the expense of intense computer computations.

5. Multiparameter Problems.

The discussion so far has centered on the simple case $\hat{\theta} \sim f_\theta$, where we have only a real-valued parameter θ and a real-valued summary statistic $\hat{\theta}$ from which we are trying to construct a confidence interval for θ. We have been able to show favorable properties of the BC_a intervals for the simple case, but of course the simple case is where we least need a general method like the bootstrap.

This section discusses the more difficult situation where there are nuisance parameters besides the parameter of interest θ. Section 6 discusses the nonparametric situation, where the number of nuisance parameters is effectively infinite. Because of the inherently simple nature of the bootstrap it will be easy to extend the BC_a method to cover these cases, though we will not be able to provide as strong a justification for the correctness of the resulting intervals.

Suppose then that the data $\underset{\sim}{y}$ comes from a parametric family $\mathcal{F}$ of density functions $f_{\underset{\sim}{\eta}}$, where $\underset{\sim}{\eta}$ is an unknown vector of parameters, and we want a confidence interval for the real-valued parameter $\theta = t(\underset{\sim}{\eta})$. In Efron (1984), the multivariate normal case $\underset{\sim}{y} \sim N_k(\underset{\sim}{\eta},\underset{\sim}{I})$ is examined in detail.

From $\boldsymbol{y}$ we obtain $\hat{\boldsymbol{\eta}}$, the MLE of $\boldsymbol{\eta}$, and $\hat{\theta} = t(\hat{\boldsymbol{\eta}})$, the MLE of θ. The BC interval for θ , (2.3), is obtained as indicated at the beginning of Section 2: step (i) of the bootstrap algorithm is to sample $\boldsymbol{y}_1^*, \boldsymbol{y}_2^* \ldots \boldsymbol{y}_B^* \overset{\text{iid}}{\sim} f_{\hat{\boldsymbol{\eta}}}$, giving $\hat{\theta}_b^* = t(\boldsymbol{y}_b^*)$, etc. However in order to obtain the BC_a intervals, we also need to know the appropriate value of a , the acceleration constant. We will find a by following Stein's (1956) construction, which replaces the multiparameter family $\mathcal{F} = \{f_{\boldsymbol{\eta}}\}$ by a *least favorable* one-parameter family $\hat{\mathcal{F}}$.

Let $\dot{\boldsymbol{\ell}}_{\boldsymbol{\eta}}$ be the vector with i^{th} coordinate $\frac{\partial}{\partial \eta_i} \log f_{\boldsymbol{\eta}}(\boldsymbol{y})$ so $\dot{\boldsymbol{\ell}}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{y}) = 0$ by definition of the MLE $\hat{\boldsymbol{\eta}}$, and let $\ddot{\boldsymbol{\ell}}_{\hat{\boldsymbol{\eta}}}$ be the $k \times k$ matrix with ij^{th} entry $\frac{\partial^2}{\partial \eta_i \partial \eta_i} \log f_{\boldsymbol{\eta}}(\boldsymbol{y})|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}}$. Also let $\hat{\boldsymbol{\nabla}}$ be the gradient vector of $\theta = t(\boldsymbol{\eta})$ evaluated at the MLE, $\hat{\nabla}_i = \frac{\partial}{\partial \eta_i} t(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}}$. The *least favorable direction* at $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$ is defined to be

$$\hat{\boldsymbol{\mu}} \equiv (-\ddot{\boldsymbol{\ell}}_{\hat{\boldsymbol{\eta}}})^{-1} \hat{\boldsymbol{\nabla}} . \tag{5.1}$$

Then the least favorable family $\hat{\mathcal{F}}$ is the one-parameter subfamily of $\mathcal{F}$ passing through $\hat{\boldsymbol{\eta}}$ in the direction $\hat{\boldsymbol{\mu}}$,

$$\hat{\mathcal{F}} = \{\hat{f}_\tau(\boldsymbol{y}^*) \equiv f_{\hat{\boldsymbol{\eta}}+\tau\hat{\boldsymbol{\mu}}}(\boldsymbol{y}^*)\} . \tag{5.2}$$

Using $\boldsymbol{y}^*$ to denote a hypothetical data vector from $\hat{f}_\tau$ is intended to avoid confusion with the actual data vector $\boldsymbol{y}$ which gave $\hat{\boldsymbol{\eta}}$; $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\mu}}$ are fixed in (5.2), only τ being unknown.

Consider the problem of estimating $\theta(\tau) \equiv t(\hat{\boldsymbol{\eta}}+\tau\hat{\boldsymbol{\mu}})$ having observed $\boldsymbol{y}^* \sim \hat{f}_\tau$. The Fisher information bound for an unbiased estimate of θ in this one-parameter family evaluated at $\tau = 0$, equals $\hat{\boldsymbol{\nabla}}'(-\ddot{\boldsymbol{\ell}}_{\hat{\boldsymbol{\eta}}})^{-1}\hat{\boldsymbol{\nabla}}$, which is the same as the corresponding bound for estimating $\theta = t(\boldsymbol{\eta})$, at $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$, in the multiparameter family $\mathcal{F}$. This is Stein's reason for calling $\hat{\mathcal{F}}$ least favorable.

We will use $\hat{F}$ to calculate an approximate value for the acceleration constant a,

$$a \doteq \frac{\text{SKEW}_{\tau=0} \dfrac{\partial \log f_{\hat{\eta}+\tau\hat{\mu}}(y^*)}{\partial\tau}}{6} . \tag{5.3}$$

This is formula (3.1) applied to $\hat{F}$, assuming that $\hat{\tau} = 0$ (which is the MLE of τ in $\hat{F}$ when $y^* = y$, the actual data vector).

Formula (5.3) is especially simple in the exponential family case where the densities $f_\eta(y)$ are of the form

$$f_\eta(y) = e^{n[\eta' y - \psi(\eta)]} f_0(y) . \tag{5.4}$$

The factor n in the exponent of (5.4) isn't necessary, but is included to agree with the situation where the data consists of i.i.d. observations $x_1, x_2, \ldots, x_n$, each with density $e^{\eta' x - \psi(\eta)}$, and y is the sufficient vector $\Sigma_{i=1}^{n} x_i/n$.

Lemma 3. For the exponential family (5.4), formula (5.3) gives

$$a = \frac{1}{6\sqrt{n}} \frac{\hat{\psi}^{(3)}(0)}{(\hat{\psi}^{(2)}(0))^{3/2}} , \tag{5.5}$$

where

$$\hat{\psi}^{(j)}(0) = \left.\frac{\partial^j \psi(\hat{\eta}+\tau\hat{\mu})}{\partial\tau^j}\right|_{\tau=0} . \tag{5.6}$$

Proof: We have

$$\left.\frac{\partial \log f_{\hat{\eta}+\tau\hat{\mu}}(y^*)}{\partial\tau}\right|_{\tau=0} = n\hat{\mu}'(y^* - \hat{\psi}(\hat{\eta})) , \tag{5.7}$$

so $\text{SKEW}_{\tau=0} \dfrac{\partial \log f_{\hat{\eta}+\tau\hat{\mu}}(y^*)}{\partial\tau}$ equals the skewness of $\hat{\mu}' y^*$ for $y^* \sim f_{\hat{\eta}}$. The fact that $\text{SKEW}(\hat{\mu}' y^*)$ equals $[\hat{\psi}^{(3)}(0)/(\hat{\psi}^{(2)}(0))^{3/2}]/\sqrt{n}$ is a standard exercise in exponential family theory. Note: Lemma 3 applies to $y \sim N_k(\eta, I)$, the case considered in Efron (1984), and gives $a = 0$, which is why the unaccelerated BC intervals worked well there.

Table 3 relates to the following example:

$$\underset{\sim}{y} \sim N_4(\underset{\sim}{\eta}, \sigma^2_{\underset{\sim}{\eta}} I) \qquad [\sigma_{\underset{\sim}{\eta}} = 1+a(\|\eta\|-8)] \ , \tag{5.8}$$

where we observe $\underset{\sim}{y} = (8,0,0,0)$ and wish to set confidence intervals for the parameter $\theta = t(\underset{\sim}{\eta}) = \|\eta\|$. The case $a = 0$ amounts to finding a confidence interval for the non-centrality parameter of a noncentral χ^2 distribution, and can be solved exactly. The theory of Efron (1984) applies to the $a = 0$ case, and we see that the BC_0 interval, i.e. the ordinary BC interval (2.3), well matches the exact interval.

	Exact	(R/L)	BC_a	(R/L)	(5.3)
$a = .10$	[6.46, 9.69]	(.96)	[6.47, 9.70]	(.97)	.0984
$a = .05$	[6.32, 9.57]	(.85)	[6.34, 9.56]	(.84)	.0498
$a = 0$	[6.14, 9.47]	(.74)	[6.19, 9.44]	(.75)	0
$a = -.05$	[5.92, 9.38]	(.65)	[6.03, 9.35]	(.66)	-.0498
$a = -.10$	[5.62, 9.30]	(.56)	[5.89, 9.27]	(.60)	-.0984

Table 3. Central 90% confidence intervals for $\theta = \|\eta\|$, having observed $\|y\| = 8$, from the parametric family $\underset{\sim}{y} \sim N_4(\underset{\sim}{\eta}, \sigma^2_{\underset{\sim}{\eta}} I)$, with $\sigma_{\eta} = 1+a(\|\eta\|-8)$. The standard interval (1.1) is [6.36,9.64] for all values of a. The last column shows that (5.3) nearly equals the constant a in this case. The exact intervals are based on the non-central χ^2 distribution.

Table 3 shows the result of varying the constant a from $.10$ to $-.10$. This example has a particularly simple geometry; the sphere $C_{\hat\theta} = \{\underset{\sim}{\eta} : \|\underset{\sim}{\eta}\| = \hat\theta\}$ is the set of $\underset{\sim}{\eta}$ vectors having $t(\underset{\sim}{\eta})$ equal to the MLE value $\hat\theta = t(\hat{\underset{\sim}{\eta}})$; the least favorable direction $\hat{\underset{\sim}{\mu}}$ is orthogonal to $C_{\hat\theta}$ at $\hat{\underset{\sim}{\eta}}$; the distribution of $\hat\theta$ is nearly normal (see Table 2 of Efron, 1984), with standard deviation changing at rate nearly equal a, as in (3.10). The BC_a intervals alter predictably with a. For instance comparing the upper endpoint at $a = .10$ with $a = 0$, notice that $(9.70-8.00)/(9.44-8.00) = 1.18$, closely matching the obvious expansion factor due to acceleration, $1 + .10 \cdot 1.654 = 1.17$.

We could disguise problem (5.8) by making non-linear transformations

$$\tilde{\underset{\sim}{y}} = g(\underset{\sim}{y}), \quad \tilde{\underset{\sim}{\eta}} = h(\underset{\sim}{\eta}) , \tag{5.9}$$

in which case the geometry of the BC_a intervals might not be obvious from the form of the parameter $\theta = t(h^{-1}(\tilde{\underset{\sim}{\eta}})) = ||h^{-1}(\tilde{\underset{\sim}{\eta}})||$ and the transformed densities $\tilde{f}_{\tilde{\underset{\sim}{\eta}}}(\underset{\sim}{y})$. However the BC_a method is invariant under such transformations, see Remark C, Section 10, so the statistician would automatically get the same intervals as if he knew the normalizing transformations $\underset{\sim}{y} = g^{-1}(\hat{\underset{\sim}{y}})$, $\underset{\sim}{\eta} = h^{-1}(\hat{\underset{\sim}{\eta}})$.

Currently we cannot justify the BC_a method as being second order correct in the multiparameter context of this section, though it seems a likely conjecture that this is so. We know that it is so in the one-parameter case, Section 4, and in the restricted multiparameter case of Efron (1984), where the BC_a and BC methods coincide; and that the BC_a method makes a rather obvious correction to the BC interval in the general multiparameter case.

6. The Nonparametric Case.

This section concerns the nonparametric case where the data $\underset{\sim}{y} = (x_1,\ldots,x_n)$ consists of i.i.d. observations x_i which may have come from any probability distribution F on their common sample space $\mathcal{X}$. There is a real-valued parameter $\theta = t(F)$ for which we desire an approximate confidence interval. We will show how the BC_a method can be used to provide such an interval based on the obvious nonparametric estimate $\hat{\theta} = t(\hat{F})$. Here $\hat{F}$ is the empirical probability distribution of the sample, putting mass $1/n$ on each observed value x_i.

The nonparametric bootstrap cdf $\hat{G}$ of $\hat{\theta}^*$ is generated by following steps (i), (ii), (iii) at the beginning of Section 2: each bootstrap data set $\underset{\sim}{y}_b^*$ equals $(x_{1b}^*, x_{2b}^*, \ldots, x_{nb}^*)$, an i.i.d. sample of size n from $\hat{F}$. This gives a bootstrap empirical distribution $\hat{F}_b^*$ putting mass $1/n$ at each x_{ib}^* , and a corresponding bootstrap estimate $\hat{\theta}_b^* = t(\hat{F}_b^*)$. The observed standard deviation of the $\hat{\theta}_b^*$ values is the nonparametric bootstrap estimate of standard error for $\hat{\theta}$, Efron (1979). In this paper we are pursuing the more difficult task of constructing approximate confidence intervals from the bootstrap distribution.

At this point we could use $\hat{G}$ to form the BC interval (2.3), but for the BC_a interval (2.6) we also need the value of a. We will derive a simple formula for a , based on Lemma 3. It depends on

$$U_i = \lim_{\Delta \to 0} \frac{t((1-\Delta)\hat{F}+\Delta\delta_i)-t(\hat{F})}{\Delta} , \quad (i = 1,2,\ldots,n) , \tag{6.1}$$

the <u>empirical influence function</u> of $\hat{\theta} = t(\hat{F})$. Here δ_i is a point mass at x_i , so U_i is the derivative of the estimate $\hat{\theta}$ with respect to the mass on point x_i. Definition (6.1) assumes that $t(F)$ is smoothly defined for choices of F near $\hat{F}$, see Section (6.3) of Efron (1982), or Section 5 of Efron (1979). [Note: $\Sigma_i^n U_i = 0$.]

The next section shows that Lemma 3, applied to a family appropriate to the nonparametric situation, gives the following approximation for the constant a

$$a \doteq \frac{1}{6} \frac{\Sigma_{i=1}^n U_i^3}{(\Sigma_{i=1}^n U_i^2)^{3/2}} . \tag{6.2}$$

This is a convenient formula since the U_i can be evaluated easily using finite difference in definition (6.1).

<u>Example 1: The law school data.</u> Table 4 shows two indices of student excellence, LSAT and GPA, for each of 15 American law schools, see Section 2.2

i	(LSAT,GPA)	U_i	i	(LSAT,GPA)	U_i
1:	(576,3.39)	-1.507	9:	(651,3.36)	0.310
2:	(635,3.30)	0.168	10:	(605,3.13)	0.004
3:	(558,2.81)	0.273	11:	(653,3.12)	-0.526
4:	(578,3.03)	0.004	12:	(575,2.74)	-0.091
5:	(666,3.44)	0.525	13:	(545,2.76)	0.434
6:	(580,3.07)	-0.049	14:	(572,2.88)	0.125
7:	(555,3.00)	-0.100	15:	(594,2.96)	-0.048
8:	(661,3.43)	0.477			

Table 4. The law school data, and values of the empirical influence function for the correlation coefficient $\hat{\rho}$.

of Efron (1982). The Pearson correlation coefficient $\hat{\rho}$ between LSAT and GPA equals .776 ; we want a confidence interval for the true correlation ρ. Table 3 also shows the values of U_i for the statistic $\hat{\rho}$, from which formula (6.2) produces $a \doteq -.0817$. $B = 100,000$ bootstrap replications (about 100 times more than actually needed, see Section 8) gave $z_0 = \Phi^{-1}(.463) = -.0927$, definition (2.2). Taking $\alpha = .05$ in (2.6), (2.7) resulted in the central 90% BC_a interval $[.43,.92]$ for ρ. The corresponding bivariate normal interval, based on Fisher's $\tanh^{-1}$ transformation, is $[.49,.90]$. The standard interval (1.1), $\hat{\rho} \pm 1.645\, \hat{\sigma}$, using the bootstrap estimate $\hat{\sigma} = .133$, is $[.56,.99]$.

Formula (4.2) is invariant under monotone changes of the parameter of interest. This results in the BC_a intervals having correct transformation properties. Suppose for example that we change parameters from ρ to $\phi = g(\rho) \equiv \tanh^{-1}(\rho)$, with corresponding nonparametric estimate $\hat{\phi} = g(\hat{\rho})$. The central 90% BC_a interval for ϕ based on $\hat{\phi}$ is then the obvious transformation of the interval for θ based on $\hat{\theta}$, $[g(.43),g(.92)] = [.46,1.59]$. This compares with Fisher's $\tanh^{-1}$ interval $[g(.49),g(.93)] = [.54,1.47]$ and the standard interval $\hat{\phi} \pm 1.645\, \hat{\sigma}_{\phi} = [.49,1.59]$. The standard interval is much more reasonable-looking on the $\tanh^{-1}$ scale, as we might expect from

Fisher's transformation theory. As commented before, a major advantage of the BC_a method is that the statistician need not know the correct scale to work on. In effect the method effectively selects the best (most normal) scale, and then transforms the interval back to the scale of interest.

Example 2: Mardia, Kent, and Bitty (1979), pages 3, and 234, give 5 test scores for each of $n = 88$ students. The principal eigenvector of the 5×5 sample covariance matrix accounts for $\hat{\theta} = .619$ of the total variation, i.e., $.619 = (\text{largest eigenvalue})/(\text{sum of eigenvalues})$. Suppose we want a central 90% confidence interval for the corresponding population parameter.

Table 5 shows the BC_a intervals based on $B = 1000$ bootstrap replications: $z_0 = .095$, $a = .0194$ (6.2), so the BC_a interval is nearly the same as the BC interval in this case. Both are nearly the same as the standard interval (1.1). (Here we have used the bootstrap standard error estimate .046 rather than the asymptotic normal-theory estimate .041.) In this case the standard interval is quite acceptable, though this is evident only after the bootstrap analysis. For a random sample of 22 of the 88 students, the standard interval agreed less well with the BC_a interval, $B = 4000$ bootstrap replications.

	All 88 Students		Random 22 Students	
BC_a	[.537, .691]	(R/L=.88)	[.550,.825]	(R/L=.75)
Standard	[.543, .695]		[.574,.847]	
MLE	.619		.711	
$(z_0,a,\hat{\sigma}_B)$	(-.095,.0194,.046)		(-.084,.0327,.083)	

Table 5. Central 90% approximate confidence intervals for the proportion of total variability due to the first principal component; test score data from Mardia, Kent, and Bibby (1979). The standard intervals are based on the bootstrap estimate of standard error.

Example 3: The mean. Suppose F is a distribution on the real line, and $\theta = t(F)$ equals the expectation $E_F X$. The empirical influence function $U_i = (x_i - \bar{x})$, so (6.2) gives

$$a = \frac{1}{6} \frac{\Sigma(x_i-\bar{x})^3}{[\Sigma(x_i-\bar{x})^2]^{3/2}} = \frac{1}{6\sqrt{n}} \frac{\hat{\mu}_3}{\hat{\mu}_2^{3/2}} = \frac{\hat{\gamma}}{6\sqrt{n}} . \qquad (6.3)$$

Here $\hat{\mu}_h = \Sigma(x_i-\bar{x})^h/n$, the h^{th} sample central moment, and $\hat{\gamma} = \hat{\mu}_3/\hat{\mu}_2^{3/2}$, the sample skewness. It turns out also that $z_0 \doteq \hat{\gamma}/6\sqrt{n}$ in this case, by standard Edgeworth arguments. Both a and z_0 are typically of order $n^{-\frac{1}{2}}$.

Because the sample mean is such a simple statistic, we can use Edgeworth methods to get asymptotic expressions for the α-level endpoint of the BC_a interval:

$$\theta_{BC_a}[\alpha] = \bar{x} + \hat{\sigma}\{z^{(\alpha)} + \frac{\hat{\gamma}}{6\sqrt{n}}(2z^{(\alpha)^2}+1) + O_p(n^{-1})\} \qquad (6.4)$$

$\hat{\sigma} \equiv (\hat{\mu}_2/n)^{\frac{1}{2}}$. This compares with

$$\theta_{BC}[\alpha] \doteq \bar{x} + \hat{\sigma}\{z^{(\alpha)} + \frac{\hat{\gamma}}{6\sqrt{n}}(z^{(\alpha)^2} + 1) + O_p(n^{-1})\} , \qquad (6.5)$$

for the BC interval (2.3), so the BC_a intervals are shifted approximately $(\hat{\gamma}/6\sqrt{n})z^{(\alpha)2}$ further right.

Johnson (1978) suggested modifying the usual t statistic $T = (\bar{x}-\theta)/\hat{\sigma}$ to $T_J = T + (\hat{\gamma}/6\sqrt{n})(2T^2+1)$, and then considering T_J to have a standard t_{n-1} distribution, in order to obtain confidence intervals for $\theta = E_f X$. Section 10 of Efron (1981) shows that this is much like using the bootstrap distribution of $T^* = (\bar{x}^*-\bar{x})/\hat{\sigma}^*$ (rather than of $\bar{x}^*-\bar{x}$) as a pivotal quantity. Interestingly enough, the Edgeworth expansion of $\theta_J[\alpha]$, the α endpoint of Johnson's interval, coincides with (6.4). The BC_a method makes a "t correction" in the case of $\theta = E_F X$, but it is not the familiar student's t correction, which operates at third order in (1.2), but rather a second-order correction, coming from the correlation between $\bar{x}$ and $\hat{\sigma}$ in non-normal populations. See Remark D, Section 10.

The author conjectures that the nonparametric BC_a intervals will be second-order correct for any parameter θ. There is no proof of this, a major difficulty being the definition of second-order correctness in the nonparametric situation. Whether or not it is true, small-sample nonparametric confidence intervals are far from well understood, and should be interpreted with some caution:

Example 4: The variance. Suppose $\mathcal{X}$ is the real line, and $\theta = \mathrm{Var}_F X$, the variance. Line 5 of Table 2 shows the result of applying the nonparametric BC_a method to data sets $x_1, x_2, \ldots, x_{20}$ which were actually i.i.d. samples from a $N(0,1)$ distribution. The number .640 for example is the average of $\theta_{BC_a}[.05]/\hat{\theta}$ over 40 such data sets, $B = 4000$ bootstrap replications per data set. The upper limit $1.68 \cdot \hat{\theta}$ is noticably small, as pointed out by Schenker (1983). The reason is simple: the nonparametric bootstrap distribution of $\hat{\theta}^*$ has a short upper tail; compared to the parametric bootstrap distribution which is a scaled χ^2_{19} random variable. The results of Beran [1984], Bickel and Friedman [1981], and Singh (1981) show that the nonparametric bootstrap distribution is highly accurate asymptotically, but of course that isn't a guarantee of good small-sample behavior. Bootstrapping from a smoothed version of F, as in Section 5.3 of Efron (1982) alleviates the problem in this particular example.

7. Geometry of the Nonparametric Case.

Formula (6.2), which allows us to apply the BC_a method nonparametrically, is based on a simple heuristic argument: instead of the actual sample-space $\mathcal{X}$ of the data points x_i, consider only distribution F supported on $\hat{\mathcal{X}} = \{x_1, x_2, \ldots, x_n\}$, the observed data set. This is an n-category multinomial family, to which we can apply the results of Section 5. Because the multinomial is an exponential family, Lemma 3 directly gives (6.2).

We will now examine this argument more carefully, with the help of a simple geometric representation. See Section 11 of Efron (1981) for further discussion of this approach to nonparametric confidence intervals.

A typical distribution supported on $\hat{X}$ is

$$F(\underset{\sim}{w}) : \text{mass } w_i \text{ on } x_i \; . \tag{7.1}$$

where $\underset{\sim}{w} = (w_1, w_2, \ldots, w_n)$ can be any vector in the simplex $\mathcal{S}_n = \{\underset{\sim}{w} : \underset{\sim}{w} > 0, \ \Sigma_1^n w_i = 1\}$. The parameter $\theta = t(F)$ is defined on $\mathcal{S}_n$ by $\theta(\underset{\sim}{w}) = t(F(\underset{\sim}{w}))$. The central point of the simplex,

$$\underset{\sim}{w}^0 \equiv \frac{\underset{\sim}{1}}{n} = (1/n, 1/n, \ldots, 1/n) \; , \tag{7.2}$$

corresponds to $F(\underset{\sim}{w}^0)$ $\hat{F}$, the usual empirical distribution ; $\theta(\underset{\sim}{w}^0) = \hat{\theta} = t(\hat{F})$, the nonparametric MLE of θ. The curved surface

$$C_{\hat{\theta}} = \{\underset{\sim}{w} : \theta(\underset{\sim}{w}) = \theta(\underset{\sim}{w}^0) = \hat{\theta}\} \tag{7.3}$$

comprises those distributions $F(\underset{\sim}{w})$ having $\theta(\underset{\sim}{w}) = \hat{\theta}$. The vector $\underset{\sim}{U}_i$ is orthogonal to $C_{\hat{\theta}}$ at $\underset{\sim}{w}^0$, as shown in Figure 1, which follows from definition (6.1) of the empirical influence function. ($\underset{\sim}{U}$ is essentially the gradient of $\theta(\underset{\sim}{w})$ at $\underset{\sim}{w}^0$, see Efron (1982), Section 6.3.)

With $\underset{\sim}{w}$ unknown, but $\hat{X} = \{x_1, \ldots, x_n\}$ considered fixed, we can imagine setting a confidence interval for $\theta(\underset{\sim}{w})$ on the basis of a hypothetical sample $x_1^*, x_2^*, \ldots, x_n^* \overset{\text{iid}}{\sim} F(\underset{\sim}{w})$. A sufficient statistic is the vector of proportions $P_i = \#\{x_j^* = x_i\}/n$, say $\underset{\sim}{P} = (P_1, P_2, \ldots, P_n)$, with distribution

$$\underset{\sim}{P} \sim \text{Mult}_n(n, \underset{\sim}{w})/n \; , \quad (\underset{\sim}{w} \in \mathcal{S}_n) \; . \tag{7.4}$$

The notation here indicates n draws from an n-category multinomial, having probability w_i for category n. We suppose that we have observed $\underset{\sim}{P} = \underset{\sim}{w}^0$ in

(7.4), i.e. that the hypothetical sample $x_1^*, \dots, x_n^*$ equals the actual sample $x_1, \dots, x_n$.

Distributions (7.4) form an n-parameter exponential family (5.4), with $\underset{\sim}{y} = \underset{\sim}{P}$, $\eta_i = \log(nw_i)+c$, and $\psi(\eta) = \log(\Sigma_1^n e^{\eta_i}/n)$. Here c can be any constant, since all vectors $\underset{\sim}{\eta} + c\underset{\sim}{1}$ correspond to the same probability vector $\underset{\sim}{w}$, namely $w_i = e^{\eta_i}/\Sigma_1^n e^{\eta_j}$.

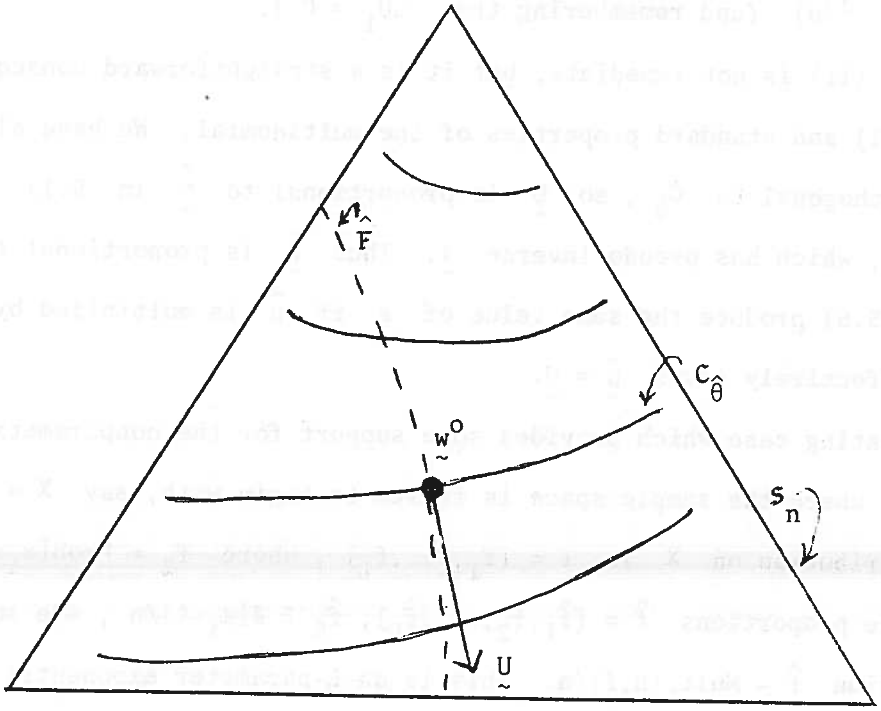


Figure 1. All probability distributions supported on $\{x_1, x_2, \dots, x_n\}$ are represented as the simplex $\mathcal{S}_n$. The central point $\underset{\sim}{w}^o$ corresponds to the empirical distribution $\hat{F}$. The curves indicate level surfaces of constant value of the parameter θ. In particular $C_{\hat{\theta}}$ comprises those probability distributions having θ equal to $\theta(\underset{\sim}{w}^o) = \hat{\theta}$, the MLE. The least favorable family $\hat{F}$ passes through $\underset{\sim}{w}^o$ in the direction $\underset{\sim}{U}$, orthogonal to $C_{\hat{\theta}}$.

If one accepts the reduction of the original nonparametric problem to (7.4), with observed value $\underset{\sim}{P} = \underset{\sim}{w}^o$, then it is easy to carry through the least favorable family calculations (5.1)-(5.2): (i) $\hat{\underset{\sim}{\eta}} = \underset{\sim}{0}$; (ii) $\hat{\underset{\sim}{\mu}} = \underset{\sim}{U}$; (iii) $\hat{f}_\tau$ is the member of (7.4) corresponding to $\hat{\underset{\sim}{\eta}}+\tau\hat{\underset{\sim}{\mu}} = \tau\underset{\sim}{U}$, namely

$$\underset{\sim}{P}^* \sim \mathrm{Mult}(n,\underset{\sim}{w}^\tau)/n \qquad (w_i^\tau = e^{\tau U_i}/\Sigma_{j=1}^n e^{\tau U_j}) \ ; \qquad (7.5)$$

(iv) finally, formula (6.2) follows directly from Lemma 3, by differentiating $\hat{\psi}(\tau) = \log(\Sigma_1^n e^{\tau U_j}/n)$ (and remembering that $\Sigma U_i = 0$.).

Only step (ii) is not immediate, but it is a straightforward consequence of definition (5.1) and standard properties of the multinomial. We have already noted that $\underset{\sim}{U}$ is orthogonal to $C_{\hat{\theta}}$, so $\underset{\sim}{U}$ is proportional to $\hat{\underset{\sim}{\gamma}}$ in (5.1). However $-\ddot{\underset{\sim}{\ell}}_{\hat{\underset{\sim}{\eta}}} = \underset{\sim}{I} - \frac{\underset{\sim}{1}\underset{\sim}{1}'}{n}$, which has pseudo-inverse $\underset{\sim}{I}$. Thus $\hat{\underset{\sim}{\mu}}$ is proportional to $\underset{\sim}{U}$. Since (5.5), (5.6) produce the same value of a if $\hat{\underset{\sim}{\mu}}$ is multiplied by any constant, this effectively gives $\hat{\underset{\sim}{\mu}} = \underset{\sim}{U}$.

An interesting case which provides some support for the nonparametric BC_a method is that where the sample space is finite to begin with, say $X = \{1,2,\ldots,L\}$. A typical distribution on X is $\underset{\sim}{f} = (f_1,\ldots,f_L)$, where $f_\ell = \mathrm{Prob}\{x_i=\ell\}$. The observed sample proportions $\hat{\underset{\sim}{f}} = (\hat{f}_1,\hat{f}_2,\ldots,\hat{f}_L)$, $\hat{f}_\ell \equiv \#\{x_i=\ell\}/n$, are sufficient, with distribution $\hat{\underset{\sim}{f}} \sim \mathrm{Mult}_L(n,\underset{\sim}{f})/n$. This is an L-parameter exponential family, so the theory of Section 5 applies. It turns out that Lemma 3 agrees with formula (6.2) in this case. Nonparametric BC_a intervals are the same as parametric BC_a intervals when X is finite. See remarks G and H of Efron (1979), for the first-order bootstrap asymptotics of finite sample spaces.

Family (7.4) was used in Section 11 of Efron (1981) to motivate a method called nonparametric tilting, a nonparametric analogue of the standard hypothesis-testing approach to confidence interval construction. The one-parameter tilting family, (11.12) of Efron (1981), is closely related to the least favorable family $\hat{F}$ in Figure 1. Table 5 of Efron (1981) considers samples of size $n = 15$ for the one-sided exponential density $f(x) = e^{-(x+1)}$, $x > -1$. Central 90% tilting intervals for $\theta = E_F X$ were constructed for each of ten such samples, averaging $[-.34,.50]$. The corresponding nonparametric BC_a intervals averaged $[-.34,.52]$, and were quite similar to the tilting intervals on a sample-by-sample comparison. The nonparametric BC_a method is computationally simpler than nonparametric tilting, and seems likely to give similar results in most problems.

We end this section with a useful approximation formula for the bias-correction constant z_0, developed jointly with Timothy Hesterberg. In addition to (6.1) we need the second order empirical influence function

$$V_{ij} = \lim_{\Delta\to 0} \frac{t((1-2\Delta)\hat{F}+\Delta\delta_i+\Delta\delta_j) - t((1-\Delta)\hat{F}+\Delta\delta_i) - t((1-\Delta)\hat{F}+\Delta\delta_j) + t(\hat{F})}{\Delta^2} \ . \qquad (7.6)$$

Define $z_{01} \equiv (1/6)\Sigma_1^n U_i^3/(\Sigma_1^n U_i^2)^{3/2}$ (the right side of (6.2)) and

$$z_{02} \equiv \left[\frac{\underset{\sim}{U}'\underset{\sim}{V}\underset{\sim}{U}}{\|\underset{\sim}{U}\|^2} - \text{tr}\ \underset{\sim}{V}\right] / (2n\|\underset{\sim}{U}\|) \ , \qquad (7.7)$$

where $\underset{\sim}{V}$ is the $n \times n$ matrix (V_{ij}).

Lemma 4. The bias-correction constant z_0 approximately equals

$$\Phi^{-1}\{2\Phi(z_{01})\Phi(z_{02})\} \ . \qquad (7.8)$$

For the law school data, Example 1, $z_{01} = -.0817$ and $z_{02} = -.0067$, giving $z_0 = -.0869$ from (7.8), compared to $z_0 = -.0927 \pm .0039$ from $B = 100{,}000$ bootstrap replications.

The term z_{01} relates to skewness in $\hat{F}$ while z_{02} is a geometrical term arising from the curvature of $C_{\hat{\theta}}$ at $\underset{\sim}{w}^o$. It is analogous to formula (A15) of Efron (1984). Lemma 4 will not be proved here, but is referred to in Section 8.

8. Bootstrap Sample Sizes.

How many bootstrap replications need we take? So far we have pretended that $B = \infty$, but if Monte Carlo methods are necessary to obtain the bootstrap distribution (2.1), then B must be finite, usually the smaller the better. This section gives rough estimates of how small B may be taken in practice. The results are presented without proof, all being standard exercises in error estimation, see for instance Chapter 10 of Kendall and Stuart (1958).

First consider the easy problem of estimating the standard error of $\hat{\theta}$ via the bootstrap. The bootstrap estimate based on B replications, $\hat{\sigma}_B = [\Sigma_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}_\cdot^*)^2/(B-1)]^{\frac{1}{2}}$, has conditional coefficient of variation (standard deviation divided by expectation)

$$CV\{\hat{\sigma}_B|\underset{\sim}{y}\} \doteq [\frac{\hat{\delta}+2}{4B}]^{\frac{1}{2}}, \qquad (8.1)$$

where $\hat{\delta}$ is the kurtosis of the bootstrap distribution $\hat{G}$. The notation indicates that the observed data $\underset{\sim}{y}$ is fixed in this calculation. As $B \to \infty$, then (8.1) $\to 0$ and $\hat{\sigma}_B \to \sigma\hat{}$, the ideal bootstrap estimate of standard error.

Of course $\hat{\sigma}$ itself will usually not estimate the true standard error $\sigma \equiv SD_\theta\{\hat{\theta}\}$ perfectly. Let $CV(\hat{\sigma})$ be the coefficient of variation of $\hat{\sigma}$, unconditional now, averaging over the possible realizations of $\underset{\sim}{y}$. [For example if $n = 20$, $\hat{\theta} = \bar{x}$, $x_i \overset{iid}{\sim} N(0,1)$, then $CV(\hat{\sigma}) \doteq (1/40)^{\frac{1}{2}} = .16$.] The unconditional CV of $\hat{\sigma}_B$ is then approximated by

$$CV(\hat{\sigma}_B) \doteq [CV^2(\hat{\sigma}) + \frac{E\hat{\delta}+2}{4B}]^{\frac{1}{2}}. \qquad (8.2)$$

Table 6 displays $CV(\hat{\sigma}_B)$ for various choices of B and $CV(\hat{\sigma})$, assuming $E\hat{\delta} = 0$. For values of $CV(\hat{\sigma}) \geq .10$, typical in practice, there is little improvement past $B = 100$. In fact B as small as 25 gives reasonable results.

B →

		25	50	100	200	∞
CV($\hat{\sigma}$)	.25	.29	.27	.26	.25	.25
↓	.20	.24	.22	.21	.21	.20
	.15	.21	.18	.17	.16	.15
	.10	.17	.14	.12	.11	.10
	.05	.15	.11	.09	.07	.05
	0	.14	.10	.07	.05	0

Table 6. Coefficient of variation of $\hat{\sigma}_B$, the bootstrap estimate of standard error, as a function of B, the number bootstrap replications, and CV($\hat{\sigma}$), the limiting CV as $B \to \infty$. Based on (8.2). assuming $E\hat{\delta} = 0$.

Now we return to bootstrap confidence intervals. Let $\theta_B[\alpha]$ be the α-endpoint of either the BC or BC_a interval obtained from B bootstrap replications (either parametrically or nonparametrically). Let $\theta[\alpha] = \lim_{B\to\infty} \theta_B[\alpha]$. The following formula for the conditional CV of $\theta_B[\alpha]-\theta[\alpha]$ assumes that the bootstrap cdf $\hat{G}$ is roughly normal, and that z_0 and a are known, for example from (7.8) and (5.3) or (6.2):

$$CV\{\theta_B[\alpha] - \theta[\alpha] | \underset{\sim}{y}\} \doteq \frac{1}{B^{\frac{1}{2}} |z^{(\alpha)}|} \left\{ \frac{\alpha(1-\alpha)}{\phi(z^{(\alpha)})^2} \right\}^{\frac{1}{2}} , \qquad (8.3)$$

$\phi(z) \equiv e^{-\frac{1}{2}z^2}/\sqrt{2\pi}$.

Here is a brief tabulation of (8.3) × $B^{\frac{1}{2}}$,

α :	.75	.90	.95	.975
(8.3)×$B^{\frac{1}{2}}$:	4.08	1.78	1.65	1.86

. (8.4)

If B = 1000 for instance, then $CV\{\theta_B[.95]-\theta[.95]|\underset{\sim}{y}\} \doteq 1.65/1000^{\frac{1}{2}} = .052$. Reducing B to 250 increases the conditional CV to .104. This last figure will often be too big. The whole purpose of developing a theory better than (1.1) is to capture second order effects. As our examples have indicated, these become interesting when the asymmetry ratio R/L is larger than say 1.25, or smaller than .75. In such borderline situations, an extra 10% error in each tail due to inadequate bootstrap sampling is unacceptable.

If the bias-correction constant z_0 is estimated by Monte Carlo directly from (2.2), rather than from (7.8), then

$$\mathrm{CV}\{\theta_B[\alpha]-\theta[\alpha]\,|\,\underset{\sim}{y}\} \doteq \frac{1}{B^{\frac{1}{2}}\, z^{(\alpha)}}\left\{\frac{1}{\phi(0)^2} - \frac{2(1-\alpha)}{\phi(0)\phi(z^{(\alpha)})} + \frac{\alpha(1-)}{\phi(z^{(\alpha)})^2}\right\}^{\frac{1}{2}} \tag{8.5}$$

for $\alpha > .50$. This gives larger CVs than (8.3),

α :	.75	.90	.95	.975
$(8.5)\times B^{\frac{1}{2}}$:	9.23	3.87	3.07	2.94

(8.6)

Comparing (8.6) with (8.4) shows that we need B to be about four times larger to get the same CV if z_0 is estimated rather than calculated. Formula (7.8) can be very helpful!

Both (8.3) and (8.5) assume that the bootstrap cdf $\hat{G}$ is estimated by straightforward Monte Carlo sampling, as in (2.1). Professor M. V. Johns (personal communication) has developed importance sampling methods which greatly accelerate the estimation of $\hat{G}$ in some situations.

9. One-Parameter Families.

We return to the simple situation $\hat{\theta} \sim f_\theta$, where there are no nuisance parameters, and where we want a confidence interval for the real-valued parameter θ based on a real-valued summary statistic $\hat{\theta}$. This section gives a more extensive discussion of the acceleration constant a, which has played a basic role in our considerations. Three familiar types of one-parameter families will be investigated: exponential families, translation families, and transformation families.

Efron (1982A) considers the following question: for a given family $\hat{\theta} \sim f_\theta$, do there exist mappings $\hat{\phi} = g(\hat{\theta})$, $\phi = h(\theta)$ such that $\hat{\phi} = \phi+\sigma_\phi q(Z)$, $Z \sim N(0,1)$, as in (3.3)? This last form, a General Scaled Transformation Family, generalizes the concept of the ideal normalization, where $\hat{\phi} = \phi+Z$.

The question above is answered in terms of the diagnostic function $D(z,\theta) \equiv [\phi(0)/\phi(z)][\dot{F}_\theta(\hat{\theta}_\theta{}^{(\alpha)})/\dot{F}_\theta(\mu_\theta)]$. Here $\phi(z)$ is the standard normal density $(2\pi)^{-\frac{1}{2}} e^{-z^2/2}$; F_θ is the cdf $F_\theta(s) = \text{Prob}_\theta\{\hat{\theta}\leq s\}$; $\dot{F}_\theta(s) = \frac{\partial}{\partial\theta} F_\theta(s)$; $\alpha = \Phi(z)$; $\hat{\theta}_\theta^{(\alpha)}$ is the 100•α percentile of $\hat{\theta}$ given θ, $\hat{\theta}_\theta^{(\alpha)} = F_\theta^{-1}(\alpha)$; and μ_θ is the median of $\hat{\theta}$ given θ, $\mu_\theta = \hat{\theta}_\theta{}^{(.5)} = \hat{F}_\theta^{-1}(.5)$. It is shown that the form of σ_ϕ and $q(z)$ in (3.3) can be inferred from $D(z,\theta)$, the main advantage being that $D(z,\theta)$ is computed without knowledge of the normalizing transformations g, h.

The connection of transformation family theory with the acceleration constant a is the following: define

$$\varepsilon_\theta \equiv \frac{\partial}{\partial z} D(z,\theta)\bigg|_{z=0} . \tag{9.1}$$

If $q(z)$ in (3.3) is symetrically distributed about zero, a situation called a Symmetric Scaled Transformation Family (SSTF), then

$$\varepsilon_\theta = \frac{d\sigma_\phi}{d\phi} , \tag{9.2}$$

see (4.11) of Efron (1982A). A more complicated relationship holds for the GSTF case.

Notice that (9.2) is quite close to our original description of "a" as the rate of change of standard deviation on the normalized scale. As a matter of fact, we can transform (2.4), (2.5) into an SSTF by considering the statistic

$$\tilde{\phi} = \hat{\phi} + \frac{z_0}{1-az_0}\sigma_{\hat{\phi}} = \hat{\phi} + \frac{z_0}{1-az_0}(1+a\hat{\phi}) , \tag{9.3}$$

instead of $\hat{\phi}$ itself. Then it is easy to show that

$$\tilde{\phi} = \phi + (1+\varepsilon_0\phi)Z \qquad (\varepsilon_0 = \frac{a}{1-az_0}) \ , \tag{9.4}$$

an SSTF with $\sigma_\phi = 1+\varepsilon_0\phi$, $\dot{\sigma}_\phi = \varepsilon_0$ for all ϕ. [The quantity ε_0 has the same definition in (9.4) as in (3.6).]

Example. For $\hat{\theta} \sim \theta\chi^2_{19}/19$ as in Table 2, $\varepsilon_\theta = .1090$ for all θ (using (9.6) below). Also $z_0 = \Phi^{-1}\,\text{Prob}\{\chi^2_{19}<19\} = .1082$. The relationship $a = \varepsilon_0/(1+\varepsilon_0 z_0)$ obtained by solving for a in (9.4) gives $a = .1077$, the value used in Table 2.

We show below that under reasonable asymptotic conditions,

$$\frac{\text{SKEW}_\theta(\dot{\ell}_\theta)}{6} \doteq \varepsilon_\theta , \tag{9.5}$$

where $\varepsilon_\theta = \frac{\partial}{\partial z} D(z,\theta)\big|_{z=0}$ as in (9.1). This last definition of ε_θ can be evaluated for any family $\hat{\theta} \sim f_\theta$, assuming only that the necessary derivatives exist. The point here is that $\text{SKEW}_\theta(\dot{\ell}_\theta)/6$ always approximates ε_θ (9.1), and in SSTF families ε_θ has the acceleration interpretation (9.2).

Now to show (9.5). It is possible to reexpress (9.1) as

$$\varepsilon_\theta = -\frac{\phi(0)}{\dot{\mu}_\theta \; f_\theta(\mu_\theta)} \; \dot{\ell}_\theta(\mu_\theta) \ , \tag{9.6}$$

where $\dot{\mu}_\theta = \frac{d}{d\theta}\mu_\theta$, the rate of change of the median μ_θ with respect to θ. For notational convenience suppose that $\theta = 0$. Instead of $\hat{\theta}$, consider the statistic $X \equiv \dot{\ell}_0(\hat{\theta})/i_0$, where i_0 equals the Fisher information $E_0\dot{\ell}_0(\theta)^2$. The parameter ε_θ is invariant under one-to-one changes of statistic, so we can evaluate the right side of (9.6) in terms of X, $\varepsilon_\theta = -\phi(0)\dot{\ell}^X_\theta(\mu^X_\theta)/\dot{\mu}^X_\theta f^X_\theta(\mu^X_\theta)$.

For $\theta = 0$, X has expectation $E_0X = 0$ and standard deviation $\sigma^X_0 = i_0^{-\frac{1}{2}}$: also $\dot{\ell}^X_0(0) = 0$, since $X = 0$ implies $\theta = 0$ is a solution of the MLE equation.

Assuming the usual asymptotic convergence properties, as in (4.4), (4.6), we have the following approximations: $\dot{\mu}_0^X \doteq 1$; $\mu_0^X \doteq -\gamma_0^X i_0^{-\frac{1}{2}}/6$; $f_0^X(\mu_0^X) \doteq \phi(0) i_0^{\frac{1}{2}}$; $\dot{\ell}_0^X(\mu_0^X) \doteq -\sqrt{i_0}\ \gamma_0^X/6$. These are derived from standard Edgeworth and Taylor series arguments, which won't be presented here. Taken together they give $\varepsilon_0 \doteq \text{SKEW}_0(\dot{\ell}_0^X)/6 = \text{SKEW}_0(\dot{\ell}_0)/6$, which is (9.5). The quantity $\text{SKEW}_0(\dot{\ell}_0)/6$ is $O(n^{-\frac{1}{2}})$, and the error of approximation in (9.5) is quite small,

$$\varepsilon_0 = \frac{\text{SKEW}_0(\dot{\ell}_0)}{6}\,[1+O(n^{-1})] \ . \tag{9.7}$$

Approximation (9.5) is particularly easy to understand in one-parameter exponential families. Suppose $x_1, x_2, \ldots, x_n$ are i.i.d. observations from such a family, with sufficient statistic $y = \bar{x}$ having density $f_\theta(y) = e^{n[\theta y - \psi(\theta)]} f_0(y)$. In this case formula (9.6) becomes

$$\varepsilon_\theta = \frac{\sigma_\theta^Y\ \phi(0)}{\dot{\mu}_\theta^Y f_\theta^Y(\mu_\theta^Y)} \left[\frac{\lambda_\theta^Y - \mu_\theta^Y}{\sigma_\theta^Y}\right] , \tag{9.8}$$

where $\lambda_\theta^Y = E_\theta\{y\}$, $\mu_\theta^Y = \text{median}_\theta\{y\}$, $\dot{\mu}_\theta^Y = \partial\mu_\theta^Y/\partial\theta$, etc. The term $[(\lambda_\theta^Y - \mu_\theta^Y)/\sigma_\theta^Y] = \gamma_\theta^Y/6[1+O(n^{-1})]$, while $\sigma_\theta^Y \phi(0)/\dot{\mu}_\theta^Y f_\theta^Y(\mu_\theta^Y) = 1 + O(n^{-1})$, both of the calculations being quite straightforward. Thus $\varepsilon_\theta = \gamma_\theta^Y/6[1+O(n^{-1})]$. Since $\dot{\ell}_\theta(y) = n[y-\lambda_\theta]$, we have $\text{SKEW}_\theta(\dot{\ell}_\theta(y)) = \text{SKEW}_\theta(y) = \gamma_\theta^Y$, verifying (9.5) for one-parameter exponential families.

Example. If $Y \sim \text{Poisson}(\theta)$, $\theta = 15$, then $\text{SKEW}_\theta(\dot{\ell}_\theta)/6 = 1/(6\cdot\theta^{\frac{1}{2}}) = .0430$. For the continued version of the Poisson family used in Efron (1982A), $\frac{\partial}{\partial z} D(z,\theta)\big|_{z=0} = .0425$ for $\theta = 15$.

Translation Families. Suppose we observe a translation family $\hat{\zeta} = \zeta + W$ as in (2.9). Express W as a function $q(Z)$ of $Z \sim N(0,1)$, for simplicity assuming $q(0) = 0$ and $q'(0) = 1$ as in Efron (1982A). Then $z_0 = \Phi^{-1}\text{Prob}\{\hat{\zeta} < \zeta\} = 0$. In this case it looks like methods based on the percentiles of the boostrap distribution must give wrong answers, since if W is long-tailed to the right then the correct interval (2.10) is long-tailed to the left, and vice-versa. However the BC_a method produces at least roughly correct intervals, as we saw in the proof of Lemma 1.

What happens is the following: for any constant A the transformation $g_A(t) \equiv (e^{At}-1)/A$ gives $\hat{\phi} = g_A(\hat{\zeta})$, $\phi = g_A(\zeta)$ and $Z_A = g_A(W)$ satisfying

$$\hat{\phi} = \phi+\sigma_\phi^A \cdot Z_A \qquad (\sigma_\phi^A = 1+A\phi) \quad . \tag{9.9}$$

The Taylor series for $W = q(Z)$ begins $W = Z+(\gamma_W/6)Z^2 +\ldots$ where $\gamma_W = \mathrm{SKEW}(W)$. Then $Z_A = Z+(\gamma_W/6)^2Z^2+(A/2)Z^2 +\ldots$.

The choice $A = a \equiv -\gamma_W/3$ results in $Z_a = Z+cZ^3 +\ldots$, the quadratic term cancelling out; Z_a is then approximately normal, so (9.9) is approximately situation (2.4), (2.5), with $z_0 = 0$, $a = -\gamma_W/3$. But we know that the BC_a intervals are correct if we can transform to situation (2.4), (2.5). An application of Lemma 2, assuming $Z_a \sim N(0,1)$, shows that $a = -\gamma_W/3 \doteq \mathrm{SKEW}(\dot{\ell}_\zeta(\hat{\zeta}))/6$ for the translation family $\hat{\zeta} = \zeta+W$, reverifying (3.1). [If $Z_a \sim N(0,1)$ in (9.9) then a must equal ε , the constant value of ε_ζ , (9.1), for the translation family $\hat{\zeta} = \zeta+W$; one can show directly that $\varepsilon \doteq -\gamma_W/3$ for such a family.]

In the example $\hat{\theta} \sim \theta\chi^2_{19}/19$, the two constants z_0 and a are nearly equal. This is no fluke:

Lemma 5. If $\hat{\theta}$ is the MLE of θ in a one-parameter problem having standard asymptotic properties (4.4) or (4.6), then $z_0 \doteq a$,

$$z_0 \equiv \Phi^{-1}\, \mathrm{Prob}_\theta\{\hat{\theta}<\theta\} = \frac{\mathrm{SKEW}_\theta(\dot{\ell}_\theta)}{6}\,[1 + O(n^{-1})] \quad . \tag{9.10}$$

Proof: We follow the notation and results of DiCiccio (1984): thus k_1,k_2,k_3 equal the first three c mulants of $\dot{\ell}_\theta$ under θ; k_{01}, k_{02}, k_{03} the first three cumulants of $\ddot{\ell}_\theta$; k_{001}, the first cumulant of $\dddot{\ell}_\theta$; and $k_{11} = \text{cov}_\theta(\dot{\ell}_\theta, \ddot{\ell}_\theta)$. (So $k_2 = i_\theta$, the Fisher information.) All cumulants are assumed to be $O(n)$. Then the relative bias of $\hat{\theta}$ is

$$b \equiv \frac{E_\theta(\hat{\theta}-\theta)}{\text{Var}_\theta(\theta)^{\frac{1}{2}}} = \frac{k_{001} - 2k_3}{6k_2^{3/2}} + O(n^{-3/2}) , \tag{9.11}$$

while $\hat{\theta}$ has skewness

$$\gamma_\theta = \frac{k_{001}-k_3}{k_2^{3/2}} + O(n^{-3/2}) . \tag{9.12}$$

Both b and γ_θ are $O(n^{-\frac{1}{2}})$.

Standard Edgeworth theory now gives

$$\begin{aligned} \text{Prob}_\theta\{\hat{\theta}<\theta\} &= \Phi(-b) - \frac{\gamma}{6}\, \phi(b)(b^2-1) + O(n^{-3/2}) \\ &= .5 + \phi(0)\, \frac{(2k_3-k_{001}) + (k_{001}-k_3)}{6k_2^{3/2}} + O(n^{-3/2}) \\ &= .5 + \phi(0)\, \frac{k_3}{6k_2^{3/2}} + O(n^{-3/2}) . \end{aligned}$$

Since $\text{SKEW}_\theta(\dot{\ell}_\theta) = k_3/k_2^{3/2}$, this verifies (9.10). □

In multiparameter problems it is no longer true that $z_0 = a$. The geometry of the level surface $C_{\hat{\theta}}$ adds another term to z_0, as in (7.8).

10. Remarks.

Remark A. Suppose that instead of (2.4), (2.5) we have $\sigma_\phi = \sigma_0(1+A\phi)$, $\sigma_0 \neq 1$. The transformations $\hat{\phi}^{(0)} \equiv \hat{\phi}/\sigma_0$, $\phi^{(0)} \equiv \phi/\sigma_0$, give $\hat{\phi}^{(0)} = \phi^{(0)} + \sigma^{(0)}_{\phi^{(0)}}(Z-z_0)$, where $\sigma^{(0)}_{\phi^{(0)}} = 1+a\phi^{(0)}$ and $a = A\sigma_0$, so we are back on form (2.4), (2.5). Notice that the derivative $d(\sigma_\phi/\sigma_0)/d(\phi/\sigma_0) = a$, as in (3.10). In a similar way we can transform (2.4), (2.5) so that $\sigma_{\phi_0} = 1$ at any point ϕ_0 ; the resulting value of a satisfies (3.10).

Remark B. Instead of using $\hat{\phi}$ to estimate ϕ in (2.4), (2.5) we might change to the estimator $\hat{\phi}^{(c)} \equiv \hat{\phi} - c\sigma_{\hat{\phi}}$, for some constant c. It turns out that we are still in situation (2.4), (2.5: $\hat{\phi}^{(c)} = \phi + \sigma_{\phi}^{(c)}(Z - z_0^{(c)})$ where

$$\sigma_{\phi}^{(c)} = 1 + a^{(c)}(\phi - \phi_0^{(c)}) \qquad (\phi_0^{(c)} = c/(1-ac)) \ , \tag{10.1}$$

and $a^{(c)} = a(1-ac)$, $z_0^{(c)} = z_0 + \phi_0^{(c)}$. The choice $c = -z_0/(1-az_0)$ gives $z_0^{(c)} = 0$, as in (9.3), (9.4). The choice $c = a$ gives approximately the MLE of ϕ. Interestingly enough, the BC_a interval for ϕ based on $\hat{\phi}^{(c)}$ is the same for all choices of c. Minor changes in the choice of estimator seem to have little effect on the BC_a intervals in general, though for computational reasons it is best not to use very biased estimators, having large values of z_0.

Remark C. Section 5 uses the MLE $\hat{\theta} = t(\hat{\underset{\sim}{\eta}})$. This has one major advantage the BC_a interval for θ , based on $\hat{\theta}$, stays the same under all multivariate transformations (5.9). Stein (1956) notes that the least favorable direction $\hat{\underset{\sim}{\mu}}$ transforms in the obvious way under (5.9), $\hat{\tilde{\underset{\sim}{\mu}}} = \hat{\underset{\sim}{D}}\hat{\underset{\sim}{\mu}}$, where $\hat{\underset{\sim}{D}}$ is the matrix with ijth element $\partial\tilde{\eta}_j/\partial\eta_i|_{\underset{\sim}{\eta}=\hat{\underset{\sim}{\eta}}}$, from which it is easy to check that formula (5.3) is invariant: the constant a is assigned the same value no matter what transformations (5.9) are applied. The bootstrap distribution $\hat{G}$ is similarly invariant, as shown in Efron (1984), and so is z_0. This implies that the BC_a intervals are invariant under transformations (5.9).

Remark D. The multiparametric theory of Section 5 gives an interesting result when applied to location-scale families, $y = (x,s)$, $\eta = (\theta,\sigma)$, and the family of densities $f_{\eta}(y)$ has the form

$$f_{\theta,\sigma}(x,s) = \frac{1}{\sigma^2} f_{01}\left(\frac{x-\theta}{\sigma}, \frac{s}{\sigma}\right) , \tag{10.2}$$

$f_{01}(x,s)$ being a known bivariate density function.

Suppose we wish to set a confidence interval for the location parameter θ on the basis of its MLE $\hat{\theta}$. Parametric bootstrap intervals are based on the distribution of $\hat{\theta}^*$ when sampling from $f_{\hat{\theta},\hat{\sigma}}(x^*,s^*)$. The BC interval essentially amounts to pretending that σ is known (and equal to $\hat{\sigma}$) in (10.2), and that we have only a location problem to deal with, rather than a location-scale problem. In contrast, the BC_a interval takes account of the fact that σ is unknown. In particular the least favorable direction $\hat{\underset{\sim}{\mu}}$, plotted in the (θ,σ) plane, is not parallel to the θ axis. It has a component in the σ direction, whose magnitude is determined by the correlation between x and s. This means that Stein's least favorable family (5.2) does not treat σ as a constant.

Table 7 relates to the following choice of $f_{01}(x,s)$:

$$x \sim \frac{\chi^2_{30}}{30} - 1 , \qquad s|x \sim (1+x)(\chi^2_{14}/14)^{\frac{1}{2}} , \tag{10.3}$$

the two χ^2 variates being independent. This is a computationally more tractable version of the problem discussed in Tables 4 and 5 of Efron (1981). Approximate central 90% intervals are given for θ, having osberved $(x,s) = (0,1)$. For any other observed (x,s) the intervals transform in the obvious way, $\theta_{xs}[\alpha]=x+s\theta_{01}[\alpha]$. Line 3 shows the exact interval, based on inverting the distribution of the pivotal quantity $T = (\hat{\theta}-\theta)/\hat{\sigma}$ for situations (10.2), (10.3).

1.	BC interval:	[-.336,.501]	(R/L) = 1.49
2.	BC_a interval:	[-.303,.603]	(R/L) = 1.99
3.	T interval:	[-.336,.670]	(R/L) = 1.99

Table 7. Central 90% intervals for θ, having observed $(x,s) = (0,1)$ from the location-scale family (10.2), (10.3). Line 3 is based on the actual distribution of the pivotal quantity $T = (\hat{\theta}-\theta)/\hat{\sigma}$. The observed MLE values are $\hat{\theta} = 0$, $\hat{\sigma} = .966$.

In this case the BC_a method makes a large "second-order t correction", as in Example 3 of Section 6, shifting the BC interval a considerable ways rightward, and acheiving the correct R/L ratio. The length of the BC_a interval is 90% the length of the T interval. This deficiency is a third-order effect, in the spirit of the familiar student's t correction. It arises from the variability of $\hat{\sigma}$ as an estimate of σ, rather than the second-order effect due to the correlation of $\hat{\sigma}$ with $\hat{\theta}$.

Remark E. Section 2 says that the family $y \sim \theta\chi_{19}^2$ can be mapped into form (2.4), (2.5). What are the appropriate mappings? It simplifies the problem to consider the equivalent family $\hat{\theta} \sim \theta(\chi_{19}^2/c_0)$ where $c_0 = 18.3337 = \text{median}(\chi_{19}^2)$. Then $\hat{\zeta} \equiv g_1(\hat{\theta})$, $\zeta \equiv g_1(\theta)$, $W \equiv g_1(\chi_{19}^2/c_0)$, give a translation family (2.9), with $\text{median}(W) = 0$, for any mapping $g_1(t) = (\log t)/c_1$. Choosing $c_1 = .3292$ results in $W = q(Z)$ having $q(0) = 0$, $q'(0) = 1$, as in Section 9's discussion of translation families.

Section 9 suggests normalizing a translation family by $g_A(t) = (e^{At}-1)/A$, a good choice for A being the constant ε_θ, (9.1), which equals .1090 for all θ in the family $\hat{\theta} \sim \theta(\chi_{19}^2/c_0)$. The combined transformation $g(t) = g_A(g_1(t))$ is $g(t) = 9.1746[t^{.3311} - 1]$. The transformed family $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$ is of form (2.4), (2.5),

$$\hat{\phi} = \phi + (1+.1090\cdot\phi)Z \qquad \left(Z = 9.1746\left[\left(\frac{\chi_{19}^2}{c_0}\right)^{.3311} - 1\right]\right) . \qquad (10.4)$$

Numerical calculations verify that Z as defined in (10.4) is very close to a standard normal variate. In fact we have automatically recovered, nearly, the Wilson-Hilferty cube root transformation, Johnson and Kotz (1970). Using (10.4), it is not difficult to show that g(t) as defined above gives approximately (2.4), (2.5) when applied to the family $\hat{\theta} \sim \theta(\chi_{19}^2/19)$ considered in Section 2, with constants z_0 and a as stated.

11. Proof of Theorem 1.

A monotonic mapping $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$ transforms the exact confidence interval in the obvious way, $\phi_{EX}[\alpha] = g(\theta_{EX}[\alpha])$, and likewise for the BC_a interval. By using such a mapping we can always make $\hat{\phi} = 0$ and the distribution of $\hat{\phi}$ given $\phi = 0$ perfectly normal. Because of (4.6), which says that the distributions of $\hat{\theta}$ are approaching normality at the usual $O(n^{-\frac{1}{2}})$ rate, the normalizing transformation g is asymptotically linear, $g(\theta) = \theta + c_2\theta^2 + c_3\theta^3 + \ldots$, $c_2 = O(n^{-\frac{1}{2}})$, $c_3 = O(n^{-1})$.

We will assume that the problem is already in the form $\hat{\theta} = 0$, with the cdf of $\hat{\theta}$ for $\theta = 0$ normal, say

$$G_0 \sim N(-z_0, 1) \ . \qquad (11.1)$$

Here $z_0 = \Phi^{-1} P_0\{\hat{\theta}<0\}$ must be included because it is not affected by any monotonic transformations $z_0 \doteq \gamma_\theta/6$ is $O(n^{-\frac{1}{2}})$ by (4.6). A simple exercise using the mean value theorem of calculus shows that if (4.7) is true in the transformed problem (11.1), then it is true in the original problem.

Assuming (4.6), $\hat{\theta} = 0$, and (11.1) we will show that the exact interval has endpoint

$$\theta_{Ex}[\alpha] \doteq \frac{z_0 + z^{(\alpha)}}{1 - \dot{\sigma}_0 z^{(\alpha)} + \dot{\beta}_0 + \frac{\dot{\gamma}_0}{6}(z^{(\alpha)^2} - 1)} + \frac{\ddot{\sigma}_0}{2}(z_0 + z^{(\alpha)})^3 \ , \qquad (11.2)$$

compared to

$$\theta_{BC_a}\alpha \doteq \frac{z_0 + z^{(\alpha)}}{1 - \dot{\sigma}_0(z_0 + z^{(\alpha)})} \qquad (11.3)$$

for the BC_a interval. In this section the symbol "$\doteq$" indicates accuracy through $O(n^{-1})$, with errors $O(n^{-3/2})$. Then

$$\frac{\theta_{BC_a}[\alpha]-\theta_{Ex}[\alpha]}{\sigma_{\hat{\theta}}} \doteq \theta_{BC_a}[\alpha]\{\dot{\sigma}_0 z_0 + \dot{\beta}_0 + \frac{\dot{\gamma}_0}{6}(z^{(\alpha)2}-1)\} - \frac{\ddot{\sigma}_0}{2}(z_0+z^{(\alpha)})^3 , \quad (11.4)$$

which is $O(n^{-1})$ as claimed in Theorem 1.

The proof of (11.2) begins by noting that (11.1) implies $\beta_0 = -z_0$, $\sigma_0 = 1$, $\gamma_0 = 0$, $\delta_0 = 0$. Then (4.6) gives

$$\begin{aligned} E_\theta\hat{\theta} = \theta+\beta_\theta &\doteq (1+\dot{\beta}_0)\theta - z_0 , \quad \sigma_\theta \doteq 1+\dot{\sigma}_0\theta+\ddot{\sigma}_0\theta^2/2 , \\ \gamma_\theta &\doteq \dot{\gamma}_0\theta , \quad\quad , \quad \delta_\theta \doteq 0 , \end{aligned} \quad (11.5)$$

for $\theta = O(1)$. The 100·α percentile of $\hat{\theta}$ given θ is

$$\begin{aligned} \hat{\theta}_\theta^{(\alpha)} &\doteq (\theta+\beta_\theta) + \sigma_\theta\{z^{(\alpha)} + \frac{\gamma_\theta}{6}(z^{(\alpha)2}-1)\} \\ &\doteq [(1+\dot{\beta}_\theta)\theta - z_0] + [1+\dot{\sigma}_0\theta + \frac{\ddot{\sigma}_0}{2}\theta^2][z^{(\alpha)} + \frac{\dot{\gamma}_0\theta}{6}(z^{(\alpha)2}-1)] , \end{aligned} \quad (11.6)$$

using a Cornish-Fisher expansion and (11.5). However the θ that has $\hat{\theta}_\theta^{(\alpha)} = 0$ is by definition $\theta_{Ex}[1-\alpha]$. Solving the lower expression in (11.6) for 0, and substituting $1-\alpha$ for α, gives (11.2).

The proof of (11.3) follows from (2.6), (2.7), and (11.1), (which says that $\hat{G} \sim N(-z_0,1)$) if we can establish that $a = \dot{\sigma}_0(1+O(n^{-1}))$. In fact we show below that

$$\varepsilon_\theta = \dot{\sigma}_0(1+O(n^{-1})) \quad \text{for} \quad \theta = O(n^{-\frac{1}{2}}) , \quad (11.7)$$

which combines with $a = \varepsilon_0/(1+\varepsilon_0 z_0) = \varepsilon_0(1+O(n^{-1}))$ to give the required result.

Formula (11.7) follows from (11.5), which gives the simpler expressions

$$E_\theta\hat{\theta} \doteq \theta - z_0, \quad \sigma_\theta \doteq 1+\dot{\sigma}_0\theta, \quad \gamma_\theta \doteq 0, \quad \delta_\theta \doteq 0 \quad (11.8)$$

for $\theta = O(n^{-\frac{1}{2}})$. The cdf of $\hat{\theta}$ given θ is calculated to be

$$G_\theta(\hat{\theta}) \doteq \phi(z_\theta)\, \dot{z}_\theta - \frac{\dot{\gamma}_0}{6}(z_\theta^2-1) \quad , \tag{11.9}$$

$z_\theta \equiv (\hat{\theta}-\theta-\beta_\theta)/\sigma_\theta$, $\dot{z}_\theta = \frac{\partial}{\partial\theta} z_\theta$. Straightforward expansions give

$$D(z^{(\alpha)},\theta) \doteq \frac{1 + \dot{\sigma}_0 z^{(\alpha)} + \dot{\beta}_0 + (\dot{\gamma}_0/6)(z^{(\alpha)2}-1)}{1 + \dot{\beta}_0 - \dot{\gamma}_0/6} \quad , \tag{11.10}$$

from which $\varepsilon_\theta = \frac{\partial}{\partial z} D(z,\theta)\big|_{z=0} \doteq \dot{\sigma}_0/(1+\dot{\beta}_0-\dot{\gamma}_0/6)$, verifying (11.7), (11.3), and the main result (11.4).

The proof that $\theta_{BC_a}[\alpha]$ also matches the Cox-McCullogh formula (4.8) is similar to the proof of Theorem 1, and won't be presented here. The main step is an expression for $\theta_{BC_a}[\alpha]$ involving Lemma 5,

$$\theta_{BC_a}[\alpha] \doteq z^{(\alpha)} + (\hat{k}_3/6\hat{k}_2^{3/2})\{z^{(\alpha)2}+1\} + (\hat{k}_3/6\hat{k}_2^{3/2})^2\{2z^{(\alpha)}+z^{(\alpha)3}\} \ . \tag{11.11}$$

12. Acknowledgement.

The author is grateful to Robert Tibshirani, Timonthy Hesterberg, and John Tukey for several useful discussions, suggestions, and references.

REFERENCES

Barndorff-Nielsen, O. E. (1984). "Confidence limits from $c|\hat{j}|\bar{L}$". Report #104, Department of Theoretical Statistics, University of Aarhus.

Bartlett, M. S. (1953). "Approximate confidence intervals". *Biometrika* **40**, 12-19.

Beran, R. (1984). "Bootstrap methods in statistics". *Jber. d. Dt. Math-Verein* **86**, 14-30.

Bickel, P. J. and Freedman, D. A. (1981). "Some asymptotic theory for the bootstrap". *Annals Stat*, **9**, 1196-1217.

Cox, D. R. (1980). "Local ancillarity". *Biometrika* **67**, 279-286.

DiCiccio, T. J. (1984). "On parameter transformations and interval estimation". Technical report, Department of Mathematical Science, McMaster University.

Efron, B. (1979). "Bootstrap methods: another look at the jackknife". *Annals Stat*. **7**, 1-26.

Efron, B. (1981). "Nonparametric standard errors and confidence intervals" (with discussion). *Canadian J. of Stat*. **9**, 139-172.

Efron, B. (1982). "The jackknife, the bootstrap, and other resampling plans". SIAM-NSF, CBMS #38.

Efron, B. (1982A). "Transformation theory: how normal is a one parameter family of distributions?" *Annals Stat*. 10, 323-339. And "Maximum likelihood and decision theory". *Annals Stat*. **10**, 340-356. (Corrigenda).

Efron, B. (1984). "Bootstrap confidence intervals for parametric problems". To appear in *Biometrika*.

Efron, B. (1984A). "Comparing non-nested linear models". To appear in *JASA*, December 1984.

Fieller, E. C. (1954). "Some problems in interval estimation". *JRSS-B* **16**, 175-183.

Hougaard, P. (1982). "Parametrizations of non-linear models". *JRSS-B* **44**, 244-252.

Johnson, N. J. (1978). "Modified t tests and confidence intervals for asymmetrical populations". *JASA* **73**, 536-544.

Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions - 2*. Houghton-Mifflin, Boston.

Kendall, M. and Stuart, A. (1958). *The Advanced Theory of Statistics*. Griffen, London.

Mardia, K., Kent, J., and Bibby, J. (1979). Multivariate Analysis. Academic Press, New York.

McCullagh, P. (1984). "Local sufficiency". Biometrika **71**, 233-244.

Singh, K. (1981). "On the asymptotic accuracy of Efron's bootstrap". Annals Stat. **9**, 1187-1195.

Stein, C. (1956). "Efficient nonparametric testing and estimation". Proc. Third Berkeley Symp. 187-196.

Tukey, J. (1949). "Standard confidence points". Memorandum Report 26, Unpublished address presented to the Institute of Mathematical Statistics.